Open World Person reidentification

Three types of loss

Idnetity loss

It treats the training process of person Re-ID as an image classification problem, i.e., each identity is a distinct class.

In the testing phase, the output of the pooling layer or embedding layer is adopted as the feature extractor.

Generally, it is easy to train and automatically mine the hard samples during the training process, as demonstrated. Several works have also investigated the SoftMax variants, such as the sphere loss and AM SoftMax.

Another simple yet effective strategy, i.e., label smoothing, is generally integrated into the standard SoftMax cross-entropy loss. Its basic idea is to avoid the model fitting to over-confident annotated labels, improving the generalizability.

Verification Loss.

It optimizes the pairwise relationship, either with a contrastive loss or binary verification loss.

Binary verification discriminates the positive and negative of an input image pair.

The verification network classifies the differential feature into into positive or negative. The verification is often combined with the identity loss to improve the performance.

Triplet loss. It treats the Re-ID model training process as a retrieval ranking problem. The basic idea is that the distance between the positive pair should be smaller than the negative pair by a pre-defined margin.The combination of triplet loss and identity loss is one of the most popular solutions for deep Re-ID model learning.

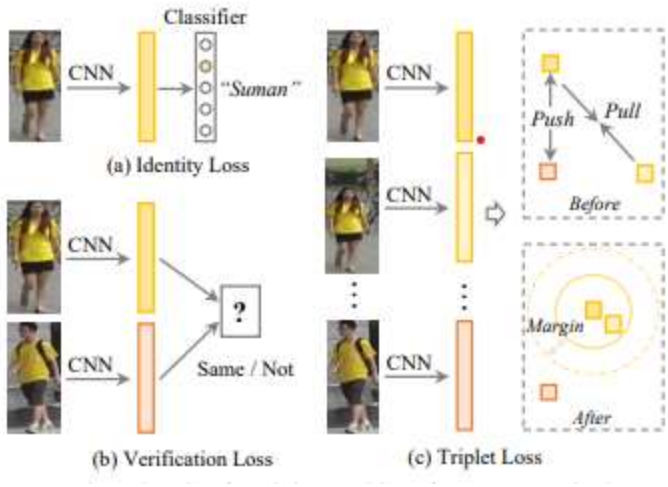


Figure: Types of loss

Process involve in person reidentification

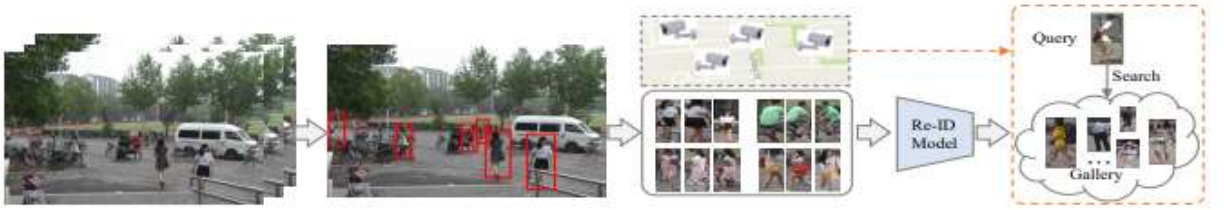


Fig: The flow of designing a practical person Re-ID system, including five main steps: 1) Raw Data Collection. (2) Bounding Box Generation. 3) Training Data Annotation. 4) Model Training and 5) Pedestrian Retrieval.

Raw Data Collection from images and videos:

I. *Access raw video data from surveillance cameras*
II. *These cameras are usually located in different places under varying environments. Most contains a large amount of complex and noisy background clutter.*

Bounding Box Generation:

i. *Extracting the bounding boxes which contain the person images from the raw video data. Generally, it is impossible to manually crop all the person images in large-scale applications.*
ii. *The bounding boxes are usually obtained by the person detection or tracking algorithms.*

Training Data Annotation:

Annotating the cross-camera labels.

Training data annotation is usually indispensable for discriminative Re-ID model learning due to the large cross-camera variations. In the existence of large domain shift , we often need to annotate the training data in every new scenario.

Step 4: Model Training:

I. *Training a discriminative and robust Re-ID model with the previous annotated person. images/videos. This step is the core for developing a Re-ID system and it is also the most widely studied paradigm in the literature.*
II. *Extensive models have been developed to handle the various challenges, concentrating on feature representation learning, distance metric learning or their combinations.*

5) Step 5: Pedestrian Retrieval:

I. *The testing phase conducts the pedestrian retrieval. Given a person-of-interest (query) and a gallery set, we extract the feature representations using the Re-ID model learned in previous stage.*
II. *A retrieved ranking list is obtained by sorting the calculated query-to-gallery similarity. Some methods have also investigated the ranking optimization to improve the retrieval performance.*

Table Closed-world vs. Open-world Person Re-ID.

Closed-world	*Open-world*
Single-modality Data	*Heterogeneous Data*
Bounding Boxes Generation	*Raw Images/Videos*
Sufficient Annotated Data	*Unavailable/Limited Labels*
Correct Annotation	*Noisy Annotation*
Query Exists in Gallery	*Open-set*

A. Heterogeneous Re-ID

A. Re-ID between depth and RGB images

Depth images capture the body shape and skeleton information. This provides the possibility for Re-ID under illumination/clothes changing environments, which is also important for personalized human interaction applications.

Research paper

"Recurrent attention models for depth-based person identification,"

Combine the convolutional and recurrent neural networks to identify small, discriminative local regions of the human body.

B. text-to-image Re-ID

Addresses the matching between a text description and RGB images

It is imperative when the visual image of query person cannot be obtained, and only a text description can be alternatively provided.

Research Paper

A gated neural attention model with recurrent neural network learns the shared features between the text description and the person images. This enables the end to-end training for text to image pedestrian retrieval.

Chenget al. (2018) propose a global discriminative image-language association learning method, capturing the identity discriminative information and local reconstructive image-language association under a reconstruction process.

A cross projection learning method (2018) also learns a shared space with image-to-text matching.

A deep adversarial graph attention convolution network (2019) is designed with graph relation mining.

C. visible-to infrared Re-ID

Visible-Infrared Re-ID handles the cross-modality matching between the daytime visible and night-time infrared images. It is important in low-lighting conditions, where the images can only be captured by infrared cameras.

Recent methods adopt the GAN technique to generate cross modality person images to reduce the cross-modality discrepancy at both image and feature level.

A dual-attentive aggregation learning method is presented to capture multi-level relations.

D. cross resolution Re-ID

Cross-Resolution Re-ID conducts the matching between low-resolution and high-resolution images, addressing the large resolution variations.

Research paper

A cascaded SR-GAN generates the high-resolution person images in a cascaded manner, incorporating the identity information.

Li etal. adopt the adversarial learning technique to obtain resolution-invariant image representations.

B.end to-end Re-ID from the raw images/videos

End-to-end Re-ID alleviates the reliance on additional step for bounding boxes generation. It involves the person Re-ID from raw images or videos, and multi-camera tracking.

Re-ID in Raw Images/Videos.

the model jointly performs the person detection and reidentification in a single framework.It is challenging due to the different focuses of two major components.

Zheng et al. present a two-stage framework, and systematically evaluate the benefits and limitations of person detection for the later stage person Re-ID.

Re-ID. Xiao et al. design an end-to-end person search system using a single convolutional neural network for joint person detection and re-identification.

A Neural Person Search Machine (NPSM) is developed to recursively refine the searching area and locate the target person by fully exploiting the contextual information between the query and the detected candidate region.

Similarly, a contextual instance expansion module is learned in a graph learning framework to improve the end-to-end person search.

A query-guided end-to-end person search system is developed using the Siamese squeeze-and-excitation network to capture the global context information with query-guided region proposal generation semi-/unsupervised learning with limited/unavailable annotated labels.

A localization refinement scheme with discriminative Re-ID feature learning is introduced to generate more reliable bounding boxes.

An Identity Discriminative Attention Reinforcement Learning (IDEAL) method selects informative regions for auto-generated bounding boxes, improving the Re-ID performance.

Yamaguchi et al. [200] investigate a more challenging problem, i.e., searching for the person from raw videos with text description. A multi-stage method with spatio-temporal person detection and multi-modal retrieval is proposed.

Multi-camera Tracking End-to-end person Re-ID is also closely related to multi-person, multi-camera tracking robust Re-ID model learning with noisy annotations.

1. *A graph-based formulation to link person hypotheses is proposed for multi-person tracking [201], where the holistic features of the full human body and body pose layout are combined as the representation for each person.*
2. *Ristani et al. learn the correlation between the multi-target multicamera tracking and person Re-ID by hard-identity mining and adaptive weighted triplet learning.*
3. *A locality aware appearance metric (LAAM) with both intra- and inter-camera relation modeling is proposed.*

C. Semi-supervised and Unsupervised Re-ID

**Unsupervised Re-ID*

cross-camera label estimation is one the popular approaches Dynamic graph matching (DGM) formulates the label estimation as a bipartite graph matching problem.

To further improve the performance, global camera network constraints are exploited for consistent matching.

Liu et al. progressively mine the labels with step-wise metric promotion.

A robust anchor embedding method iteratively assigns labels to the unlabelled tracklets to enlarge the anchor video sequences set. With the estimated labels, deep learning can be applied to learn Re-ID models.

For end-to-end unsupervised Re-ID, an iterative clustering and Re-ID model learning is presented in [205]. Similarly, the relations among samples are utilized in a hierarchical clustering framework [208].

Soft multi-label learning mines the soft label information from a reference set for unsupervised learning.

A Tracklet Association

Unsupervised Deep Learning (TAUDL) framework jointly conducts the within-camera tracklet association and model the cross-camera tracklet correlation.

Similarly, an unsupervised camera-aware similarity consistency mining method is also presented in a coarse-to-fine consistency learning scheme.

The intra-camera mining and inter-camera association is applied in a graph association framework

The semantic attributes are also adopted in Transferable Joint Attribute-Identity Deep Learning (TJAIDL) framework]. However, it is still challenging for model updating with newly arriving unlabeled data.

A Patch Net is designed to learn discriminative patch features by mining patch level similarity.

A Self-similarity Grouping (SSG) approach iteratively conducts grouping (exploits both the global body and local parts similarity for pseudo labeling) and Re-ID model training in a self-paced manner.

Semi-/Weakly supervised Re-ID. With limited label information, a one-shot metric learning method is proposed , which incorporates a deep texture representation and a color metric.

A stepwise one-shot learning method (EUG) is proposed for video-based Re-ID, gradually selecting a few candidates from unlabeled tracklets to enrich the labeled tracklet set.

A multiple instance attention learning framework uses the video-level labels for representation learning, alleviating the reliance on full annotation.

** Unsupervised Domain Adaptation Unsupervised domain adaptation (UDA)*

(UDA) transfers the knowledge on a labeled source dataset to the unlabeled target dataset.

Due to the large domain shift and powerful supervision in source dataset, it is another popular approach for unsupervised Re-ID without target dataset labels.

Target Image Generation.

Using GAN generation to transfer the source domain images to target-domain style is a popular approach for UDA Re-ID. With the generated images, this enables supervised Re-ID model learning in the unlabeled target domain.

Wei et al. [44] propose a Person Transfer Generative Adversarial Network (PTGAN), transferring the knowledge from one labeled source dataset to the unlabeled target dataset.

Preserved self-similarity and domain-dissimilarity is trained with a similarity preserving generative adversarial network (SPGAN).

A Hetero Homogeneous Learning (HHL) method simultaneously considers the camera invariance with homogeneous learning and domain connectedness with heterogeneous learning.

An adaptive transfer network decomposes the adaptation process into certain imaging factors, including illumination, resolution, camera view, etc. This strategy improves the cross-dataset performance.

Huang et al. try to suppress the background shift to minimize the domainshift problem.

Chen et al. design an instance-guided context rendering scheme to transfer the person identities.from source domain into diverse contexts in the target domain. Besides, a pose disentanglement scheme is added to improve the image generation .

A mutual mean-teacher learning scheme is also developed . However, the scalability and stability of the image generation for practical large-scale changing environment are still challenging.

Bak et al. [125] generate a synthetic dataset with different illumination conditions to model realistic indoor and outdoor lighting. The synthesized dataset increases generalizability of the learned model and can be easily adapted to a new dataset without additional supervision.

Target Domain Supervision Mining.

Some methods directly mine the supervision on the unlabeled target dataset with a well trained model from source dataset.

An exemplar memory learning scheme [106] considers three invariant cues as the supervision, including exemplar-invariance,camera invariance and neighborhood-invariance.

TheDomain-Invariant Mapping Network (DIMN) formulates a meta-learning pipeline for the domain transfer task,and a subset of source domain is sampled at each training episode to update the memory bank, enhancing the scalability and discriminability.

The camera view information is also applied as the supervision signal to reduce the domain gap.

A self-training method with progressive augmentation jointly captures the local structure and

global data distribution on the target dataset.

Recently, a self-paced contrastive learning framework with hybrid memory is developed with great success, which dynamically generates multi-level supervision signals.

The spatio-temporal information is also utilized as the supervision in TFusion . TFusion transfers the spatiotemporal patterns learned in the source domain to the target domain with a Bayesian fusion model.

Similarly, QueryAdaptive Convolution (QAConv) is developed to improve cross-dataset accuracy.

**State-of-The-Arts for Unsupervised Re-ID*

Unsupervised Re-ID has achieved increasing attention in recent years, evidenced by the increasing number of publications in top venues.

We review the SOTA for unsupervised deeply learned methods on two widely-used image-based Re-ID datasets.

First, the unsupervised Re-ID performance has increased significantly over the years. The Rank-1 accuracy/mAP increases from 54.5%/26.3% (CAMEL) to 90.3%/76.7% (SpCL) on the Market-1501 dataset within three years.

The performance for DukeMTMC dataset increases from 30.0%/16.4% to 82.9%/68.8%. The gap between the supervised upper bound and the unsupervised learning is narrowed significantly. This demonstrates the success of unsupervised Re-ID with deep learning.

Second, current unsupervised Re-ID is still underdeveloped and it can be further improved in the following

aspects:

1) The powerful attention scheme in supervised ReID methods has rarely been applied in unsupervised ReID.

2) Target domain image generation has been proved effective in some methods, but they are not applied in two best methods (PAST], SSG .

3) Using the annotated source data in the training process of the target domain is beneficial for cross-dataset learning, but it is also not included in above two methods. These observations provide the potential basis for further improvements.

Third, there is still a large gap between the unsupervised and supervised Re-ID.

For example, the rank-1 accuracy of supervised ConsAtt has achieved 96.1% on the Market1501 dataset, while the highest accuracy of unsupervised SpCL is about 90.3%.

Recently, He et al. [229] have demonstrated that unsupervised learning with large-scale unlabeled training data has the ability to outperform the supervised learning on various tasks .

D.Noise-Robust Re-ID

Re-ID usually suffers from unavoidable noise due to data collection and annotation difficulty.

We review noise-robust Re-ID from three aspects: Partial Re-ID with heavy occlusion, Re-ID with sample noise caused by detection or tracking errors, and Re-ID with label noise caused by annotation error.

Partial Re-ID.

Solve Re-ID problem with heavy occlusions, i.e., only part of the human body is visible . A fully convolutional network is adopted to generate fix-sized spatial feature maps for the incomplete person images.

Deep Spatial feature Reconstruction (DSR) is further incorporated to avoid explicit alignment by exploiting the reconstructing error.

Sun et al. [67] design a Visibility-aware Part Model (VPM) to extract sharable region-level features, thus suppressing the spatial misalignment in the incomplete images.

A foreground-aware pyramid reconstruction scheme also tries to learn from the unoccluded regions.

The Pose-Guided Feature Alignment (PGFA) exploits the pose landmarks to mine discriminative part information from occlusion noise. However, it is still challenging due to the severe partial misalignment, npredictable visible regions and distracting unshared body regions.

Meanwhile, how to adaptively adjust the matching model for different queries still needs further investigation.

Re-ID with Sample Noise.

This refers to the problem of the person images or the video sequence containing outlying regions/frames, either caused by poor detection/inaccurate tracking results.

To handle the outlying regions or background clutter within the person image, pose estimation cues or attention cues are exploited.

The basic idea is to suppress the contribution of the noisy regions in the final holistic representation.

For video sequences, set-level feature learning or frame level re-weighting are the commonly used approaches to reduce the impact of noisy frames.

Hou et al. [20] also utilize multiple video frames to auto-complete occluded regions. It is expected that more domain-specific sample noise handling designs in the future.

Re-ID with Label Noise.

Usually unavoidable due to annotation error.

Zheng et al. adopt a label smoothing technique to avoid label overfiting issues

A Distribution Net (DNet) that models the feature uncertainty is proposed in [235] for robust Re-ID model learning against label noise, reducing the impact of samples with high feature uncertainty.

Different from the general classification problem, robust Re-ID model learning suffers from limited training samples for each identity .

In addition, the unknown new identities increase additional difficulty for the robust Re-ID model learning.

Open-set Re-ID and Beyond

Open-set Re-ID is usually

The verification usually requires a learned condition τ , i.e., sim(query, gallery) > τ . Early researches design handcrafted systems .

For deep learning methods, an Adversarial PersonNet (APN) is proposed , which jointly learns a GAN module and the Re-ID feature extractor. The basic idea of this GAN is to generate realistic target like images (imposters) and enforce the feature extractor is robust to the generated image attack. Modeling feature uncertainty is also investigated in .

However, it remains quite challenging to achieve a high true target recognition and maintain low false target recognition rate .

Group Re-ID. It aims at associating the persons in groups rather than individuals . Early researches mainly focus on group representation extraction with sparse dictionary learning or covariance descriptor aggregation [240].

The multi-grain information is integrated to fully capture the characteristics of a group. Recently, the graph convolutional network is applied in [242], representing the group as a graph. The group similarity is also applied in the end-to-end person search [196] and the individual re-identification to improve the accuracy.

However, group Re-ID is still challenging since the group variation is more complicated than the individuals.

Dynamic Multi-Camera Network.

Dynamic updated multi-camera network is another challenging issue which needs model adaptation for new cameras or probes.

A human in-the-loop incremental learning method is introduced to update the Re-ID model, adapting the representation for different probe galleries.

Early research also applies the active learning for continuous Re-ID in multi-camera network.

A continuous adaptation method based on sparse non-redundant representative selection is introduced.

A transitive inference algorithm is designed to exploit the best source camera model based on a geodesic flow kernel.

Multiple environmental constraints (e.g., Camera Topology) in dense crowds and social relationships are integrated for an open-world person Re-ID system .The model adaptation and environmental factors of cameras are crucial in practical dynamic multicamera network.

Moreover, how to apply the deep learning technique for the dynamic multi-camera network is still less investigated.

Research paper

Attention Deep Model with Multi-Scale Deep Supervision for Person Re-Identification (2021)

PReID methods have used attention or multi-scale feature learning modules to enhance the discrimination of the learned deep features. However, the attention mechanisms may lose some important feature information. Moreover, the multi-scale models usually embed the multi-scale feature learning module into the backbone network, which increases the complexity of testing network. To address the two issues, we propose a multi-scale deep supervision with attention feature learning deep model for PReID. Specifically, we introduce a reverse attention module to remedy the feature information losing issue caused by the attention module, and a multi-scale feature learning layer with deep supervision to train the network.

Cross-view similarity exploration for unsupervised cross-domain person re-identification (2021)

Cross-view similarity exploration (CVSE) method, which combines style-transferred samples to optimize the CNN model and the relationship between samples.

In stage-I, use starGAN to train a style transfer model, which generates images of multiple camera styles for increasing the quantity and diversity of samples.

In stage-II,propse incremental optimization learning, which iterates between similarity grouping and CNN model optimization to progressively explore the potential similarities of all training samples.

With the purpose of reducing the impact of label noise on performance, propose a new ranking-guided triplet loss, which is on the basis of similarity and does not require any label to select reliable triple samples.

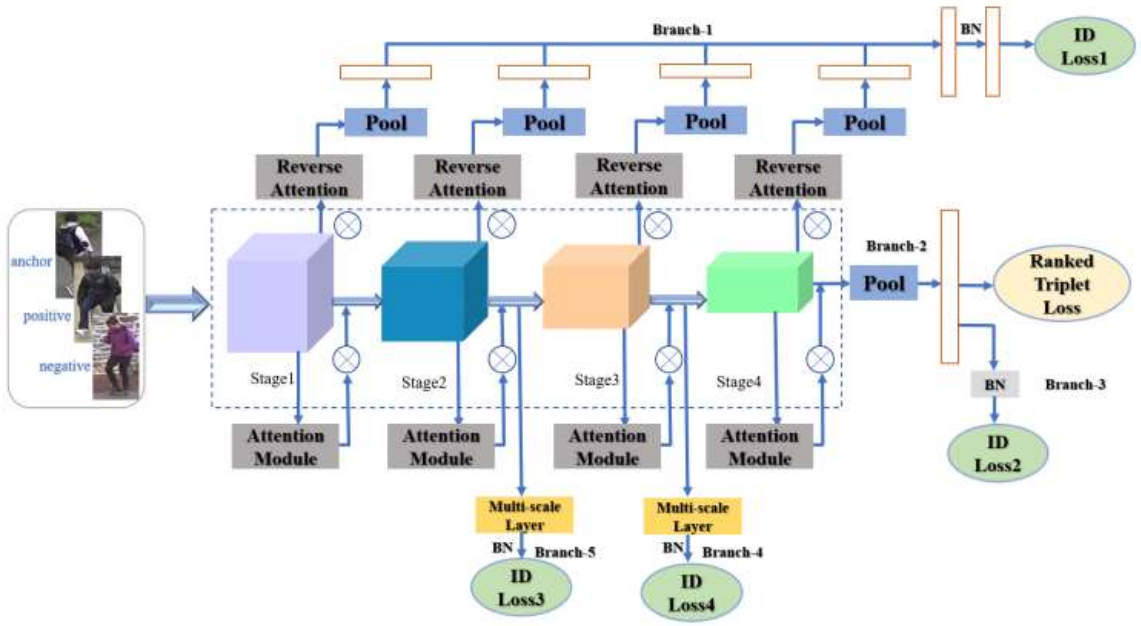


Fig: Architecture of the proposed model

we use ResNet-50 as its backbone network. The architecture consists of five branches.

Branch-1 with reverse attention block learns the feature information lost by attention block. By using the triplet and classification losses, branch-2 and branch-3 learn the global descriptors, respectively. Deep supervision with multi-scale feature learning is performed by branch-4 and branch-5. The model is supervised by four classification loss functions and one triplet loss function.

Harmonious attention network for person re-identification via complementarity between groups and individuals(2020)

A harmonious attention network for person re-identification, in which we jointly consider the complementarity between person groups and individuals.

Concretely, first we propose a two-stream attentive network (TSAN) to respectively learn the information from the person groups and individuals.

TSAN consists of a spatial–temporal fusion network for the group Re-ID, as well as a deep network for the traditionally individual person Re-ID.

To jointly consider the contributions of the groups and individuals, then propose a novel re-ranking algorithm (GIRK) based on the learned features to associate the group and individual information.

We also propose a new group Re-ID dataset DukeGroupVid to evaluate the performance of our approach. Comprehensive experimental results on the proposed dataset and other Re-ID datasets demonstrate the effectiveness of our model.

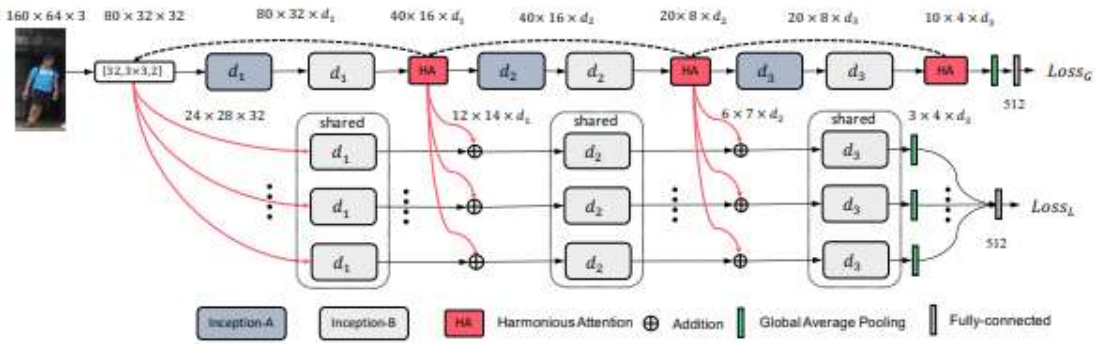


Fig: The Harmonious Attention Convolutional Neural Network. The symbol dl (l ∈ {1, 2, 3}) denotes the number of convolutional filter in the corresponding Inception unit at the l-th block

Cross-view kernel collaborative representation classification for person re-identification(2021)

Currently, person re-identification (re-ID) has been applied in many public security applications. Yet owing to the big visual appearance changes of the same identity under different views, re-ID still faces many challenges.

To reduce the intra-person discrepancy, extracting more power feature representations from pedestrian images is a reasonable solution, propose a cross-view kernel collaborative representation based classification (CV-KCRC) method for person re-ID in our work.

Our method aims to find more robust and discriminative feature representations that embody cross-view information to enhance the identification capability of features.

We map the image features into a high dimensional feature space first and then use view-specific projection matrices to project the high dimensional features into a common low dimensional subspace.

We expect that in the shared subspace the codings of same person from different views have the highest similarity and better performance can be achieved. Experiments on seven commonly used datasets reveal that our algorithm outperforms many state-of-the-art algorithms.

Learning Person Re-Identification Models From Videos With Weak Supervision(2021)

Most person re-identification methods, being supervised techniques, suffer from the burden of massive annotation requirement.

Unsupervised methods overcome this need for labeled data, but perform poorly compared to the supervised alternatives

. In order to cope with this issue, we introduce the problem of learning person re-identification models from videos with weak supervision. The weak nature of the supervision arises from the requirement of video-level labels, i.e. person identities who appear in the video, in contrast to the more precise frame-level annotations. Towards this goal, we propose a multiple instance attention learning framework for person re-identification using such video-level labels. Specifically, we first cast the video person re-identification task into a multiple instance learning setting, in which person images in a video are collected into a bag. The relations between videos with similar labels can be utilized to identify persons, on top of that, we introduce a co-person attention mechanism which mines the similarity correlations between videos with person identities in common. The attention weights are obtained based on all person images instead of person tracklets in a video, making our learned model less affected by noisy annotations. Extensive experiments demonstrate the superiority of the proposed method over the related methods on two weakly labeled person re-identification datasets.

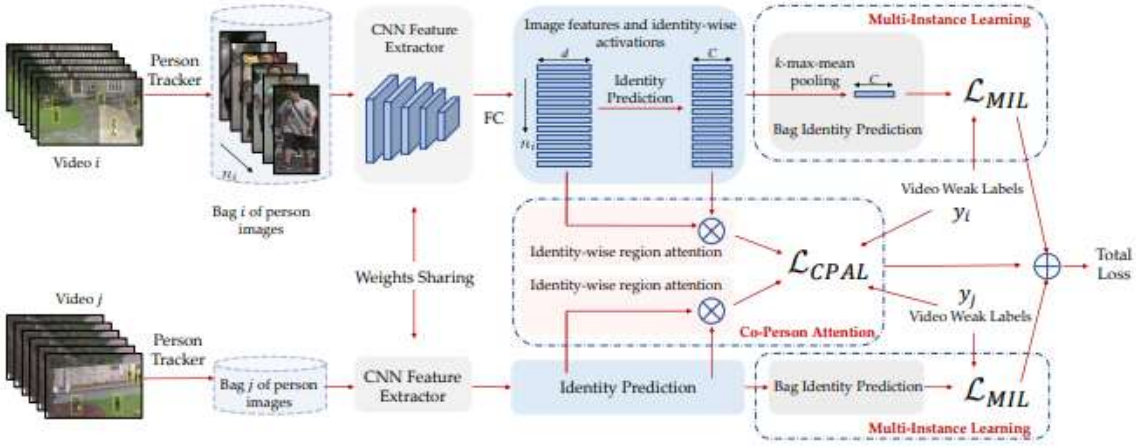


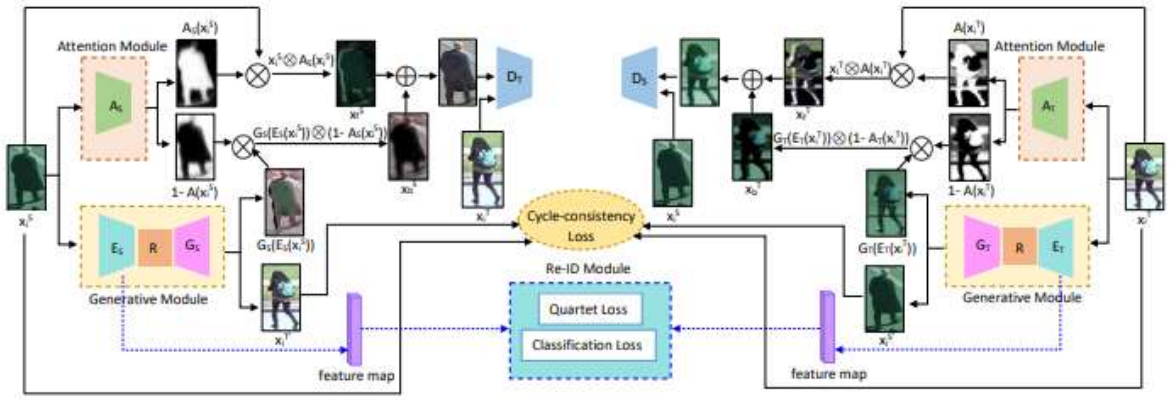
Fig: A brief illustration of our proposed multiple instance attention learning framework for video person re-id with weak supervision.

For each video, we group all person images obtained by pedestrian detection and tracking algorithms in a bag and use it as the inputs of our framework. The bags are passed through a backbone CNN to extract features for each person image. Furthermore, a fully connected (FC) layer and an identity projection layer are used to obtain identity-wise activations. On top of that, the MIL loss based on k-max-mean-pooling strategy is calculated for each video. For a pair of videos (i, j) with common person identities, we compute the CPAL loss by using high and low attention region for the common identity. Finally, the model is optimized by jointly minimizing the two loss functions.

End-to-End Domain Adaptive Attention Network for Cross-Domain Person Re-Identification

Person re-identification (re-ID) remains challenging in a real-world scenario, as it requires a trained network to generalise to totally unseen target data in the presence of variations across domains.

Recently, generative adversarial models have been widely adopted to enhance the diversity of training data. These approaches, however, often fail to generalise to other domains, as existing generative person re-identification models have a disconnect between the generative component and the discriminative feature learning stage. To address the on-going challenges regarding model generalisation, we propose an end-to-end domain adaptive attention network to jointly translate images between domains and learn discriminative re-id features in a single framework. To address the domain gap challenge, we introduce an attention module for image translation from source to target domains without affecting the identity of a person. More specifically, attention is directed to the background instead of the entire image of the person, ensuring identifying characteristics of the subject are preserved. The proposed joint learning network results in a significant performance improvement over state-of-the-art methods on several challenging benchmark datasets.



An illustration of the proposed EDAAN for image translation from one domain to another. A, E, G, and D represent the attention network, the generative module for image translation and the discriminator. A(xi) and 1 – A(xi) represent the attention maps produced by the attention network. ⊗ denotes element-wise multiplication.

<u>Cross-view kernel collaborative representation classification for person re-identification(2021)</u>

Previous person re-identification (Re-ID) methods usually focus on extracted features to against the appearance variations of pedestrians under different circumstances.

The problem caused by the scale variations did not attract much attention and is not well addressed either. In this work, we propose a novel Multi-scale Deep Feature Learning with correlation metric (MDFLCM) model to handle the scale problem in Re-ID.

Specifically, multi-scale high-level features are extracted by a specially designed end-to-end multi-scale deep convolutional network (MS-DCN) at various resolution levels.

By adding an extra correlation layer in our MDFLCM model, we can achieve the accuracy of image patch matching up to pixel-wise level.

Different from other methods extracting multi-scale features through multiple networks, we extract multi-scale features via a single network with one input image.

Extensive comparative evaluations with state-of-the-art methods on four public datasets: CUHK01, CUHK03, Market 1501 and DukeMTMC-reID, demonstrate the effectiveness of the proposed MDFLCM model on Re-ID.

Unsupervised Multi-Source Domain Adaptation for Person Re-Identification(2021)

Unsupervised domain adaptation (UDA) methods for person re-identification (re-ID) aim at transferring re-ID knowledge from labeled source data to unlabeled target data.

Among these methods, the pseudo-label-based branch has achieved great success, whereas most of them only use limited data from a single-source domain for model pre-training, making the rich labeled data insufficiently exploited.

To make full use of the valuable labeled data, we introduce the multi-source concept into UDA person re-ID field, where multiple source datasets are used during training. However, because of domain gaps, simply combining different datasets only brings limited improvement.

Address this problem from two perspectives, i.e. domain-specific view and domain-fusion view. Two constructive modules are proposed, and they are compatible with each other.

First, a rectification domain-specific batch normalization (RDSBN) module is explored to simultaneously reduce domain-specific characteristics and increase the distinctiveness of person features.

Second, a graph convolutional network (GCN) based multi-domain information fusion (MDIF) module is developed, which minimizes domain distances by fusing features of different domains. The proposed method outperforms state-of-the-art UDA person re-ID methods by a large margin, and even achieves comparable performance to the supervised approaches without any post-processing techniques.

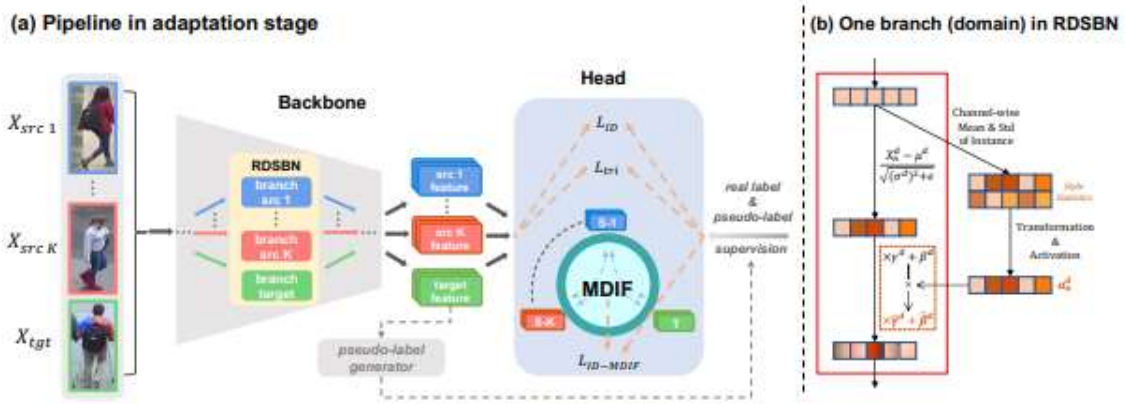


Fig: (a) The illustration of the proposed framework, including pseudo-label generator, backbone equipped with RDSBN and head equipped with MDIF. (b) The rectification operation in RDSBN. Best viewed in color.

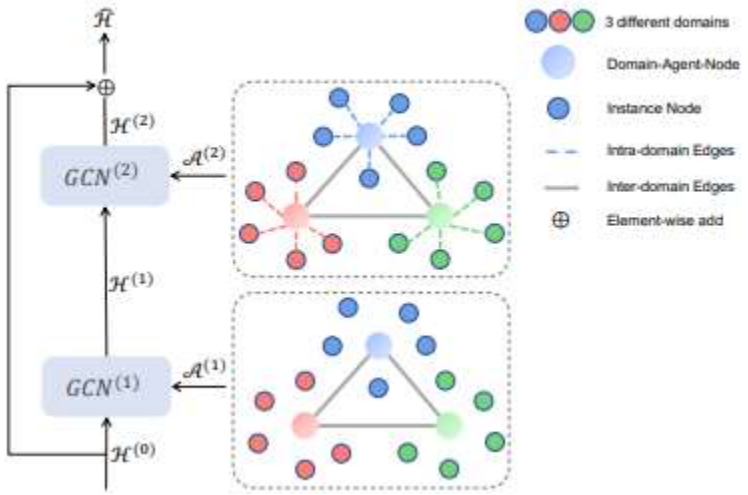


Fig: The illustration of GCN-based Multi-Domain Information Fusion module. For convenience, we present two source domains and one target domain here. Best viewed in color.

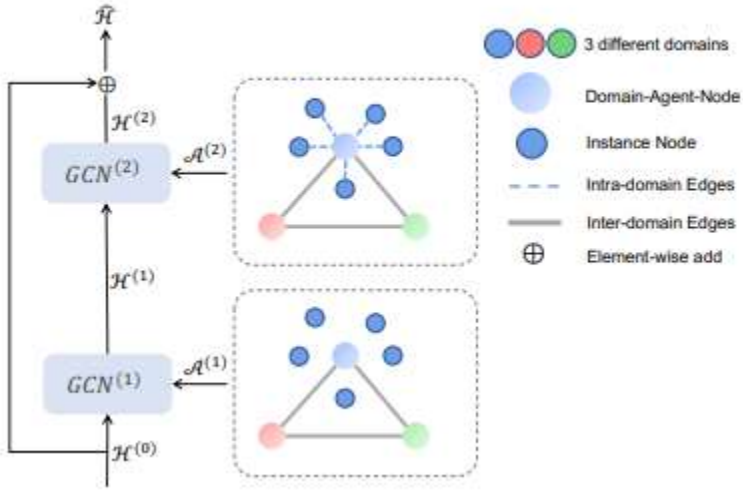


Fig: The illustration of GCN-based MDIF module in inference mode. For simplicity, we present two source domains (red and green) and one target domain (blue) here.

PMT-Net: Progressive Multi-Task Network for one-shot Person Re-Identification (2021)

Most existing person Re-Identification (Re-ID) algorithms require abundant labeled data from paired non-overlapping camera views in the fully supervised scenario.

However, the fully supervised Re-ID suffers from the limited availability of labeled training samples due to the sharply increased cost of manual efforts.

To tackle this problem, a novel Progressive Multi-Task Network (PMT-Net) for person Re-ID is proposed.

PMT-Net initializes a model using only one labeled sample for each identity, and it iteratively optimizes the model by sampling the most reliable pseudo labels dynamically from unlabeled samples.

Firstly, pedestrian attributes recognition is incorporated as an auxiliary task to learn discriminative features. Then, based on the discriminative features, the identity label for unlabeled samples is estimated by the distance between the labeled samples and unlabeled samples in feature space.

In addition, to enhance the accuracy of label estimation for the unlabeled samples, a semi-supervised clustering method, named Distance Ranked Weight Clustering (DRW-Clustering) is designed.

The clustering method weights partial unlabeled samples by the indexed ordinal of distance sorting, so that it can find the real cluster center quickly and effectively.

Extensive comparative evaluation experiments are conducted on Market1501 and DukeMTMC-reID datasets, and the experimental results indicate that the proposed method achieves performance competitive or better than that of the state-of-the-art for one-shot person Re-ID.

Appearance feature enhancement for person re-identification (2020)

A. *Person re-identification (Re-ID) has important practical application value in intelligent video analysis. Due to the illumination, occlusion, and pose variation, person Re-ID is still a challenging problem.*
B. *Some recent Re-ID methods based on ResNet-50 have achieved high accuracy, but performance degradation is caused by pose variation.*
C. *To address this issue, Pose-Invariant Convolutional Baseline (PICB) embed with the proposed Pooling Fusion Block (PFB) is put forward as a new baseline for person Re-ID task.*
D. *On the basis of PICB, an end-to-end network named Appearance-Enhanced Feature Learning Network (AEFLN) is proposed to simultaneously learn diversity body features and discriminative part features.*
E. *Specially, a novel (DBFL) strategy is presented to learn diversity body features, which could alleviate the potential local minima problem generated by optimizing model with randomly initialized parameters in PFB.*
F. *In addition, uniform part-level feature extractors are applied to learn part features, which compensates for body features' lack of distinguishable local information.*
G. *In testing phase, body features and part features are integrated to represent the enhanced appearance feature for each person image.*
H. *Comprehensive experiments have demonstrated that our method can outperform the sate-of-the-art results on several public available datasets, including Market-1501, CUHK03 and DukeMTMC-reID.*

I. *For instance, we achieve 74.8% (+11.1%) and 76.5% (+19.0%) in Rank-1 accuracy and mAP on CUHK03 dataset.*

A feature disentangling approach for person re-identification via self-supervised data augmentation(2020)

A. *To address the problem of insufficient training data in person ReID, this paper proposes a data augmentation method based on image channels shuffling, by which a large volume of diversified training samples sharing similar edges can be produced.*
B. *In the meantime, a soft label assignment strategy is designed to characterize the correlations between the original image and the generated counterparts.*
C. *Furthermore, we design an encoder–decoder based learning structure for the person ReID task, where the encoder module tackles feature disentangling according to the introduced correlations, and the decoder module handles reconstruction using the combinations of decoupled features.*
D. *Extensive experiments on four benchmark datasets demonstrate the effectiveness and robustness of the proposed method by attaining significant improvement over some state-of-the-art approaches.*

Bidirectional Interaction Network for Person Re-Identification

1. *Person re-identification (ReID) task aims to retrieve the same person across multiple spatially disjoint camera views.*
2. *Due to huge image changes caused by various factors such as posture variation and illumination transformation, images of different persons may share the more similar appearances than images of the same one.*
3. *Learning discriminative representations to distinguish details of different persons is significant for person ReID.*
4. *Many existing methods learn discriminative representations resorting to a human body part location branch which requires cumbersome expert human annotations or complex network designs.*
5. *In this article, a novel bidirectional interaction network is proposed to explore discriminative representations for person ReID without any human body part detection.*
6. *The proposed method regards multiple convolutional features as responses to various body part properties and exploits the inter-layer interaction to mine discriminative representations for person identities.*
7. *Firstly, an inter-layer bilinear pooling strategy is proposed to feasibly exploit the pairwise feature relations between two convolution layers.*
8. *Secondly, to explore interaction of multiple layers, an effective bidirectional integration strategy consisting of two different multi-layer interaction processes is designed to aggregate bilinear pooling interaction of multiple convolution layers.*
9. *The interaction of multiple layers is implemented in a layer-by-layer nesting policy to ensure the two interaction processes are different and complementary.*
10. *Extensive experiments validate the superiority of the proposed method on four popular person ReID datasets including Market-1501, DukeMTMC-ReID, CUHK03-NP and MSMT17.*

11. *Specifically, the proposed method achieves a rank-1 accuracy of 95.1% and 88.2% on Market-1501 and DukeMTMC-ReID, respectively.*

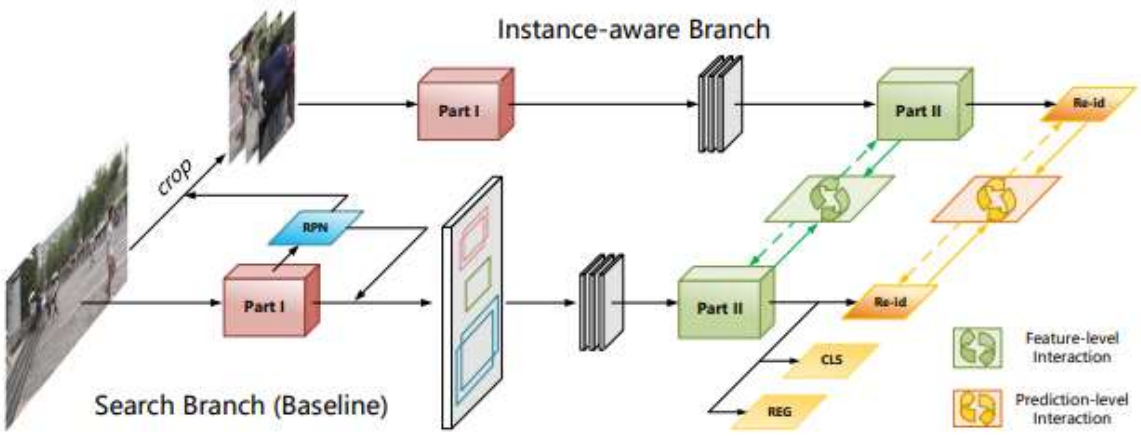


Fig: Our proposed framework.

BINet takes as inputs scene images and cropped person patches. The common parts of the two branches share parameters. Bi-directional interaction between two branches is achieved by the interaction losses. During inference, we only apply the search branch. The dashed lines represent the directions of the gradients

MEMF: Multi-level-attention embedding and multi-layer-feature fusion model for person re-identification(2021)

Person re-identification (re-ID) methods need to extract representative, rich and discriminative features in order to deal with the effect of imperfect pedestrian detectors, illumination changes, occlusions, and background confusion.

In this paper, a multi-level-attention embedding and multi-layer-feature fusion (MEMF) model is proposed for person re-ID.

Specifically, a novel backbone network is designed, in which multi-level-attention blocks are embedded into a multi-layer-feature fusion architecture.

Multi-level-attention blocks can highlight representative features and assist global feature expression, and multi-layer-feature fusion can increase the fine granularity of feature expression and obtain richer features.

Besides, a new eigenvalue difference orthogonality (EDO) loss is designed to reduce the correlation between features.

The final loss is defined as the combination of the cross-entropy loss and the EDO loss, which improves re-ID results.

The proposed method is evaluated on four popular and challenging datasets.

Detailed experiments demonstrate that the application of various elements of the MEMF model can help improve person re-ID performance. Compared with start-of-the-art methods, the MEMF model gets a promising result.

Gait recognition for person re-identification(2020)

I. *Person re-identification across multiple cameras is an essential task in computer vision applications, particularly tracking the same person in different scenes.*
II. *Gait recognition, which is the recognition based on the walking style, is mostly used for this purpose due to that human gait has unique characteristics that allow recognizing a person from a distance.*
III. *However, human recognition via gait technique could be limited with the position of captured images or videos.*
IV. *Hence, this paper proposes a gait recognition approach for person re-identification. The proposed approach starts with estimating the angle of the gait first, and this is then followed with the recognition process, which is performed using convolutional neural networks.*
V. *Herein, multitask convolutional neural network models and extracted gait energy images (GEIs) are used to estimate the angle and recognize the gait.*
VI. *GEIs are extracted by first detecting the moving objects, using background subtraction techniques. Training and testing phases are applied to the following three recognized datasets: CASIA-(B), OU-ISIR, and OU-MVLP.*
VII. *The proposed method is evaluated for background modeling using the Scene Background Modeling and Initialization (SBI) dataset.*
VIII. *The proposed gait recognition method showed an accuracy of more than 98% for almost all datasets.*
IX. *Results of the proposed approach showed higher accuracy compared to obtained results of other methods result for CASIA-(B) and OU-MVLP and form the best results for the OU-ISIR dataset.*

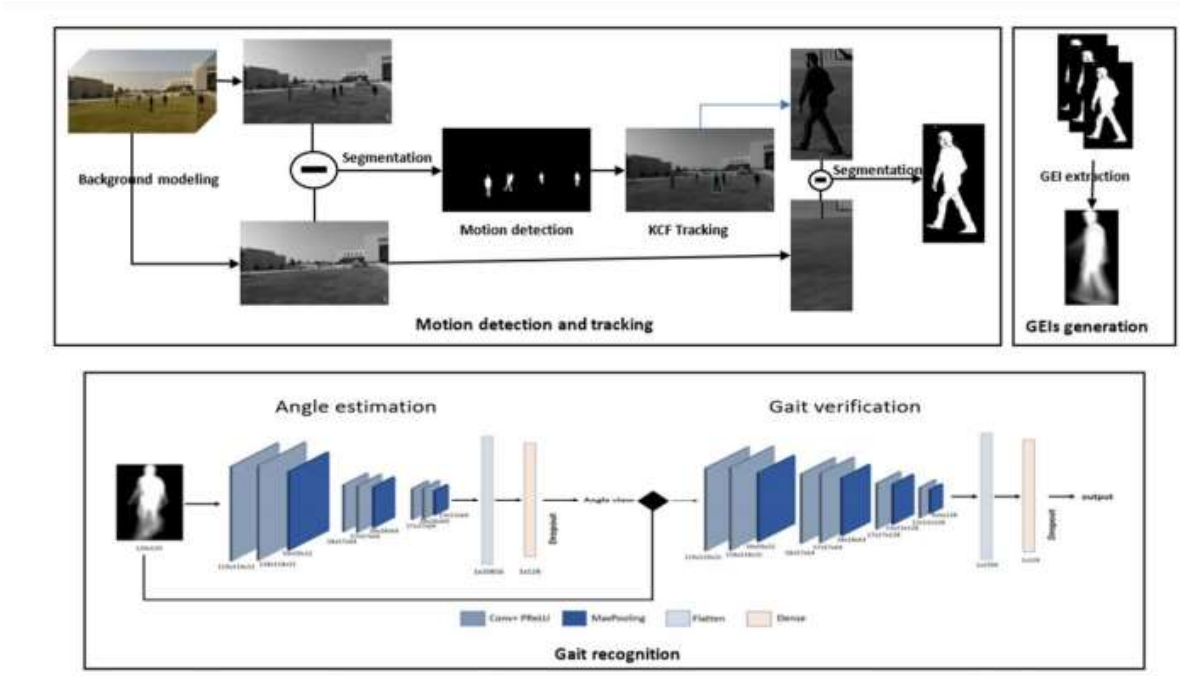


Fig: Gait recognition for person re-identification

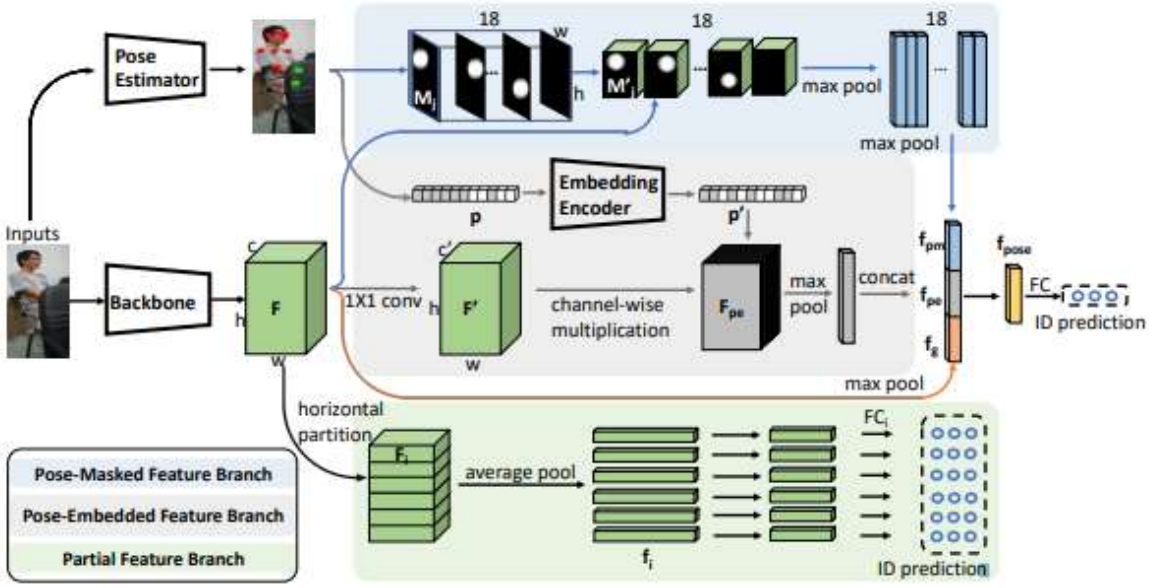
Identifying Visible Parts via Pose Estimation for Occluded Person Re-Identification(2021)

I. *We focus on the occlusion problem in person re-identification (re-id), which is one of the main challenges in real-world person retrieval scenarios.*
II. *Previous methods on the occluded re-id problem usually assume that only the probes are occluded, thereby removing occlusions by manually cropping. However, this may not always hold in practice.*
III. *This article relaxes this assumption and investigates a more general occlusion problem, where both the probe and gallery images could be occluded.*
IV. *The key to this challenging problem is depressing the noise information by identifying bodies and occlusions.*
V. *We propose to incorporate the pose information into the re-id framework, which benefits the model in three aspects.*
VI. *First, it provides the location of the body. We then design a Pose-Masked Feature Branch to make our model focus on the body region only and filter those noise features brought by occlusions.*
VII. *Second, the estimated pose reveals which body parts are visible, giving us a hint to construct more informative person features. We propose a Pose-Embedded Feature Branch to adaptively re-calibrate channel-wise feature responses based on the visible body parts.*
VIII. *Third, in testing, the estimated pose indicates which regions are informative and reliable for both probe and gallery images.*

IX. *Then we explicitly split the extracted spatial feature into parts. Only part features from those commonly visible parts are utilized in the retrieval.*

X. *To better evaluate the performances of the occluded re-id, we also propose a large-scale data set for the occluded re-id with more than 35 000 images, namely Occluded-DukeMTMC.*

XI. *Extensive experiments show our approach surpasses previous methods on the occluded, partial, and non-occluded re-id data sets.*



The pipeline of the proposed method. Red and green points indicate visible and invisible landmarks, respectively. Our model contains three branches.

In Pose-Masked Feature Branch, we generate Gaussian maps to filter out occlusions.

In the Pose-Embedded Feature Branch, we obtain the pose-embedding by the visible landmark vector.

These embeddings are further used to generate channel gates, which control the response of channels by the channel-wise multiplication.

In Partial Feature Branch, we uniformly split the extracted feature map into parts for generating part features

Spatial-Aware GAN for Unsupervised Person Re-identification(2021)

a) *The recent person re-identification research has achieved great success by learning from a large number of labeled person images.*

b) *On the other hand, the learned models often experience significant performance drops when applied to images collected in a different environment.*

c) *Unsupervised domain adaptation (UDA) has been investigated to mitigate this constraint, but most existing systems adapt images at pixel level only and ignore obvious discrepancies at spatial level.*

d) *This paper presents an innovative UDA-based person re-identification network that is capable of adapting images at both spatial and pixel levels simultaneously.*
e) *A novel disentangled cycle-consistency loss is designed which guides the learning of spatial-level and pixel-level adaptation in a collaborative manner.*
f) *In addition, a novel multi-modal mechanism is incorporated which is capable of generating images of different geometry views and augmenting training images effectively.*
g) *Extensive experiments over a number of public datasets show that the proposed UDA network achieves superior person re-identification performance as compared with the state-of-the-art.*

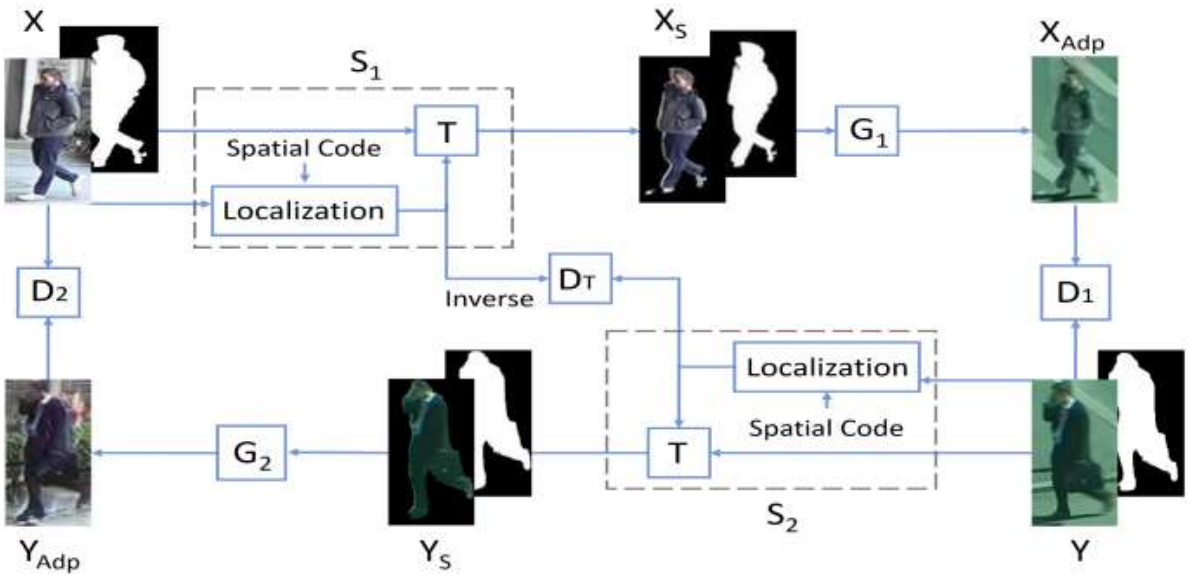


Figure: The structure of the proposed Spatial-Aware GAN (SA-GAN):

The parts within the dashed lines are two Spatial Transformer Modules S1 and S2 each of which consists of a spatial transformation T and a parameter localization network Localization.

G1, G2, D1, D2 and DT denote the generators and discriminators, respectively. X and Y denote two image domains (the binary masks are determined by U-Net), where (Xs, Ys) and (Xadp, Yadp) denote the two-domain images after the proposed spatial-level and further pixel-level adaptation, respectively

Adaptive super-resolution for person re-identification with low-resolution images(2020)

A. *Person re-identification is challenging with low-resolution query and high-resolution gallery images.*
B. *To address the resolution mismatch, many methods perform super-resolution (SR) on low-resolution queries with specifying a single scale factor.*
C. *However, using a single SR module, whichever scale factor is specified, always brings both advantages and drawbacks in recovering and identifying identity information.*

D. *A larger scale factor recovers more details but produces excessive artifacts, while a smaller one is on the contrary.*
E. *To exploit their complementary property for more robust recovery and identification, we propose the Adaptive Person Super-Resolution (APSR) model. APSR jointly trains and fuses multiple SR modules based on their generated visual contents, to fully compensate and learn the complementary identity features in an end-to-end manner.*
F. *To improve the robustness to artifacts during fusion, our model further learns informative features by online dividing and integrating the generated body regions. Extensive experiments verify the effectiveness of our method with state-of-the-art performances*

A divide-and-unite deep network for person re-identification(2020)

A. *Person re-identification (person re-ID) is one of the most challenging tasks in the field of computer vision as it involves large variations in human appearances, human poses, background illuminations, camera views, etc.*
B. *In recent literature, using part-level features for the person re-ID task provides fine-grained information, and has been proven to be effective.*
C. *Instead of relying on additional skeleton key points or pose estimation models, this paper proposes a Divide-and-Unite Network to obtain feature embedding end-to-end.*
D. *We design a deep network guided by image contents, which divides pedestrians into parts and obtains the part features with different contributions.*
E. *These part features and the global feature are united to obtain the pedestrian descriptor for person re-ID. To summarize, the contributions of this work are two-fold.*
F. *Firstly, a novel architecture of discriminative descriptor learning is proposed, which is based on the global feature and supplemented by part features.*
G. *Secondly, a Feature Division Network is constructed to generate the part features with different contributions, where the divided parts maintain the consistency of content between different images.*
H. *Extensive experiments are conducted on three widely-used benchmarks including Market1501, CUHK03, and DukeMTMC-reID. The results have demonstrated that the proposed model can achieve remarkable performance against numerous state-of-the-arts.*

Rethinking data collection for person re-identification: active redundancy reduction(2021)

A. *Annotating a large-scale image dataset is very tedious, yet necessary for training person re-identification (re-ID) models. To alleviate such a problem, we present an active redundancy reduction (ARR) framework via training an effective re-ID model with the least labeling efforts.*
B. *The proposed ARR framework actively selects informative and diverse samples for annotation by estimating their uncertainty and intra-diversity, thus it can significantly reduce the annotation workload.*
C. *Moreover, we propose a computer-assisted identity recommendation module embedded in the ARR framework to help human annotators to rapidly and accurately label the selected samples.*

D. *Extensive experiments were carried out on several public re-ID datasets to demonstrate the existence of data redundancy. Experimental results indicate that our method can reduce 57%, 63%, and 49% annotation efforts on the Market1501, MSMT17, and CUHK03, respectively, while maximizing the performance of the re-ID model.*

Fine-Grained Shape-Appearance Mutual Learning for Cloth-Changing Person Re-Identification(2021)

Recently, person re-identification (Re-ID) has achieved great progress. However, current methods largely depend on color appearance, which is not reliable when a person changes the clothes.

Cloth-changing Re-ID is challenging since pedestrian images with clothes change exhibit largen intra-class variation and small inter-class variation. Some significant features for identification are embedded in unobvious body shape differences across pedestrians.

To explore such body shape cues for cloth-changing Re-ID, we propose a Fine-grained Shape-Appearance Mutual learning framework (FSAM), a two-stream framework that learns finegrained discriminative body shape knowledge in a shape stream and transfers it to an appearance stream to complement the cloth-unrelated knowledge in the appearance features.

Specifically, in the shape stream, FSAM learns fine-grained discriminative mask with the guidance of identities and extracts fine-grained body shape features by a pose-specific multi-branch network. To complement clothunrelated shape knowledge in the appearance stream, dense interactive mutual learning is performed across low-level and high-level features to transfer knowledge from shape nstream to appearance stream, which enables the appearance stream to be deployed independently without extra computation for mask estimation.

We evaluated our method on benchmark cloth-changing Re-ID datasets and achieved the start-of-the-art performance.

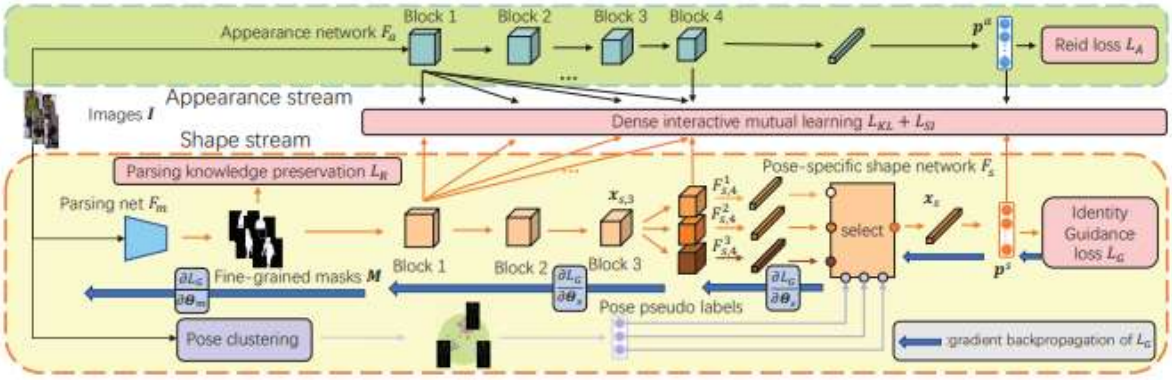


Figure: Overview of our Fine-grained Shape-Appearance Mutual learning Framework.

Our framework consists of two streams: an appearance stream and a shape stream. In the shape stream, the parsing net estimates fine-grained masks from input color images. and then the masks are fed into

the shape feature extraction network to extract fine-grained body shape features. Then, dense interactive mutual learning transfers knowledge between appearance stream and shape stream to complement fine-grained body shape feature in appearance feature. During inference, only the appearance stream is required, which saves computation cost of mask estimation.

PrGCN: Probability prediction with graph convolutional network for person re-identification(2021)

Robust similarity measurement is an important issue for person re-identification (ReID). Most existing ReID models estimate the similarity between query and gallery images by computing their Euclidean distances while ignoring the rich context information contained in the image space.

In this paper, we propose a graph convolutional network (GCN) based method to improve the similarity measurement in ReID, which regards the ReID task as a prediction problem of the link probability between node pairs.

Our method is named as PrGCN (Probability GCN), in which each person is regarded as an instance node. Firstly, an Instance Centered Sub-graphs (ICS) is constructed for each instance node to depict its rich local context information.

Secondly, the constructed ICS is input to a GCN to infer and predict the link probability of node pairs, followed by a similarity ranking between the query and gallery images according to the predicted probabilities.

Extensive experiments show that the proposed method improves the mAP and Top-1 accuracy of ReID significantly, yielding better or comparable results to the state-of-the-art methods on various benchmarks (Market1501, DukeMTMC-ReID and CUHK03).

In addition, we validate that the proposed PrGCN can be easily embedded into other deep learning architectures to replace Euclidean distance metric and achieve significant performance improvements

Dual attention-based method for occluded person re-identification(2020)

A. *Occlusion is unavoidable in real-world applications of person re-identification (ReID).*
B. *To alleviate the occlusion problem, this work proposes the detection of the occluded and visible regions of the human body by suppressing the occluded region during feature generation and matching, and enhancing the significance of the visible region.*
C. *This paper introduces a novel method based on pose-guided spatial attention (PGSA) and activation-based attention (AA) called dual-attention re-identification (DAReID).*
D. *DAReID consists of a mask branch and a global branch and uses ResNet-50 as the backbone network. The mask branch uses PGSA to obtain the visible and occluded regions of a person and constructs pose guided coarse labels for the occluded region through keypoints of the human body, driving the network to obtain robust local features.*
E. *The global branch obtains the visual activation levels of different regions through AA, and combines this with human pose information to define weighted local distances(WLD).*
F. *The WLD learning strategy is applied to drive the network to learn new and more discriminative local features.*

G. *Experimental results show that DAReID achieves comparable performance on the Market1501, DukeMTMC-reID, and CUHK-03 datasets. And on the Occluded-DukeMTMC dataset, DAReID outperforms the existing methods.*

Beyond Triplet Loss: Person Re-identification with Fine-grained Difference-aware Pairwise Loss(2020)

A. *Person Re-IDentification (ReID) aims at reidentifying persons from different viewpoints across multiple cameras.*
B. *Capturing the fine-grained appearance differences is often the key to accurate person ReID, because many identities can be differentiated only when looking into these fine grained differences.*
C. *However, most state-of-the-art person ReID approaches, typically driven by a triplet loss, fail to effectively learn the fine-grained features as they are focused more on differentiating large appearance differences.*
D. *To address this issue,we introduce a novel pairwise loss function that enables ReID models to learn the fine-grained features by adaptively enforcing an exponential penalization on the images of small differences and a bounded penalization on the images of large differences.*
E. *The proposed loss is generic and can be used as a plugin to replace the triplet loss to significantly enhance different types of state of-the-art approaches.*
F. *Experimental results on four benchmark datasets show that the proposed loss substantially outperforms a number of popular loss functions by large margins; and it also enables significantly improved data efficiency.*

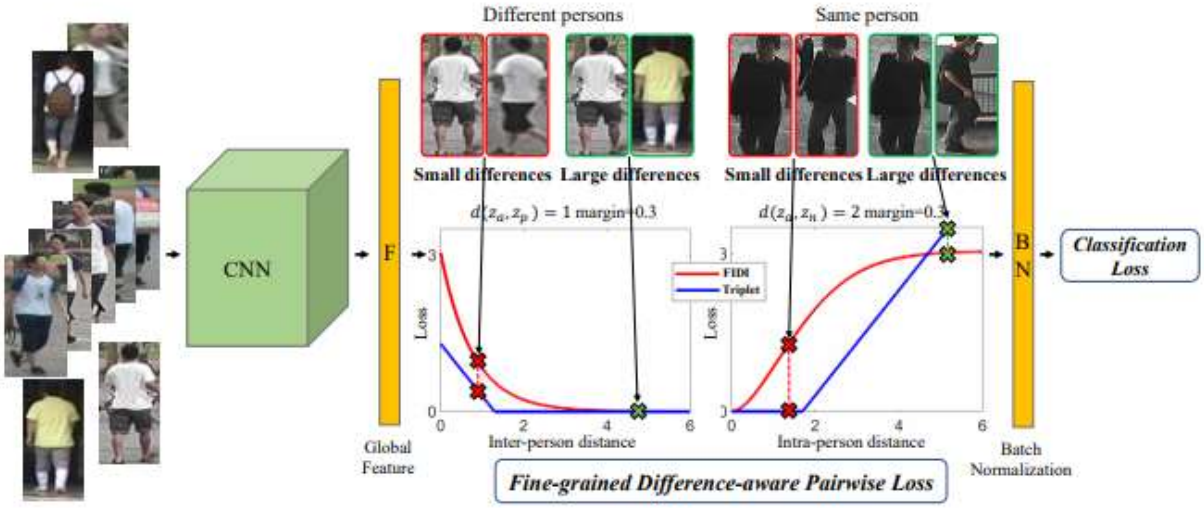


Fig: An overview of the fine-grained difference-aware pairwise loss-based framework.

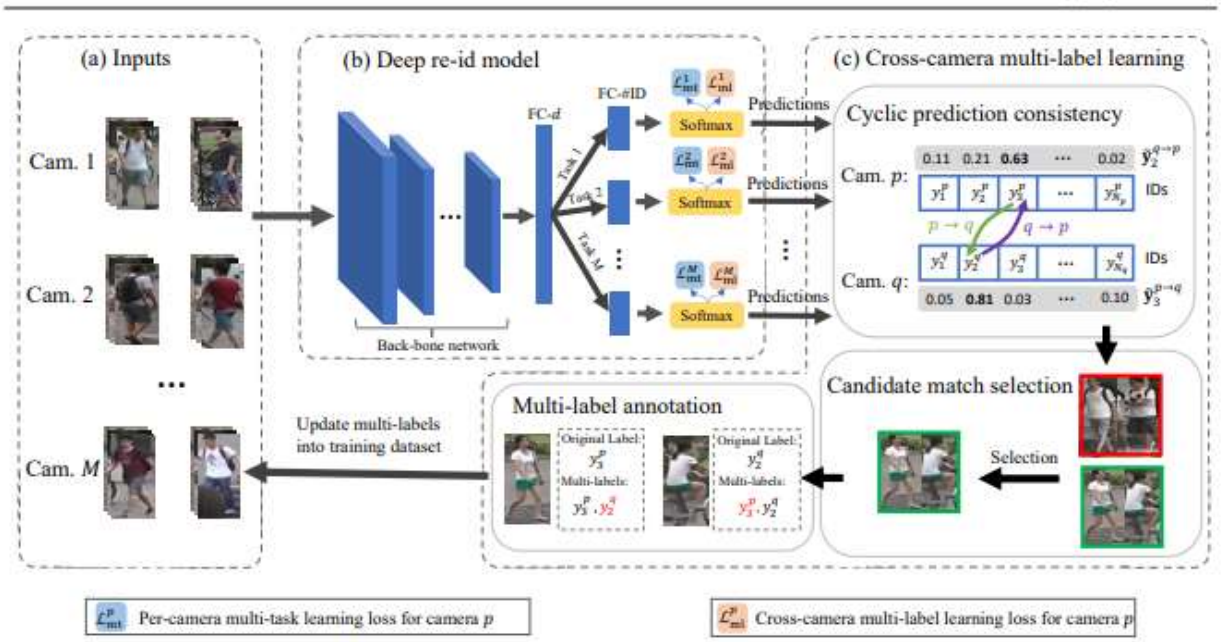
It consists of a deep CNN-based network backbone, our proposed FIDI loss and a classification loss. This framework is exactly the same as the widely-used triplet loss-based framework except that the triplet loss is replaced with our FIDI loss. The network backbone can be various CNN architectures. Unlike the triplet loss that neglects small appearance differences due to the potential dominance of unbounded

penalization on images of large intra-person differences, our FIDI loss can effectively capture the fine-grained intrapersonal/inter-person appearance differences, e.g., image pairs having small appearance difference as in the red boxes above.

We achieve this by enforcing exponentially large penalization on images of small differences and bounded penalization on images of large differences.

Intra-Camera Supervised Person Re-Identification(2021)

I. *Existing person re-identification (re-id) methods mostly exploit a large set of cross-camera identity labelled training data. This requires a tedious data collection and annotation process, leading to poor scalability in practical re-id applications.*
II. *On the other hand unsupervised re-id methods do not need identity label information, but they usually suffer from much inferior and insufficient model performance. To overcome these fundamental limitations, we propose a novel person re-identification paradigm based on an idea of independent per-camera identity annotation. This eliminates the most time-consuming and tedious inter-camera identity labelling process, significantly reducing the amount of human annotation efforts. Consequently, it gives rise to a more scalable and more feasible setting, which we call Intra-Camera Supervised (ICS) person re-id, for which we formulate a Multi-tAsk mulTi-labEl (MATE) deep learning method.*
III. *Specifically, MATE is designed for self-discovering the cross-camera identity correspondence in a per-camera multi-task inference framework. Extensive experiments demonstrate the cost-effectiveness superiority of our method over the alternative approaches on three large person re-id datasets.*
IV. *For example, MATE yields 88.7% rank-1 score on Market-1501 in the proposed ICS person re-id setting, significantly outperforming unsupervised learning models and closely approaching conventional fully supervised learning competitors.*



Overview of the proposed Multi-tAsk mulTi-labEl (MATE) deep learning method.

(a) Given per-camera independently labelled training images, MATE aims to learn an identity discriminative feature representation model. This is achieved by designing two learning components:

b) Per-camera multi-task learning where we consider each individual camera view as a separate learning task with its own identity class space and optimise these camera-specific tasks on a common feature Representation

, and (c) Cross-camera multi-task learning where we self-discover the underlying identity matching relationships across camera views via curriculum cyclic association and design a multi-label optimization algorithm to exploit this discovered cross-camera association information during model training.

The two components are integrated together ina single MATE formulation, resulting in an end-to-end trainable model.

<u>Discriminative feature extraction for video person re-identification via multi-task network(2020)</u>

1. *The goal of video-based person re-identification is to match different pedestrians in various image sequences across non-overlapping cameras. A critical issue of this task is how to exploit the useful information provided by videos. To solve this problem, we propose a novel feature learning framework for video-based person re-identification.*
2. *The proposed method aims at capturing the most significant information in the spatial and temporal domains and then building a discriminative and robust feature representation for each sequence.*

3. *More specifically, to learn more effective frame-wise features, we apply several attributes to the video-based task and build a multi-task network for the identity and attribute classifications.*
4. *In the training phase, we present a multi-loss function to minimize intra-class variances and maximize inter-class differences. After that, the feature aggregation network is employed to aggregate frame-wise features and extract the temporal information from the video.*
5. *Furthermore, considering that adjacent frames typically have similar appearance features, we propose the concept of "non-redundant appearance feature extraction" to obtain the sequence-level appearance descriptors of pedestrians. Based on the complementarity between the temporal feature and the non-redundant appearance feature, we combine them in the distance learning phase by assigning them different distance-weighted coefficients. Extensive experiments are conducted on three video-based datasets and the results demonstrate the superiority and effectiveness of our method.*

Joint Noise-Tolerant Learning and Meta Camera Shift Adaptation for Unsupervised Person Re-Identification(2021)

i. *This paper considers the problem of unsupervised person re-identification (re-ID), which aims to learn discriminative models with unlabeled data.*
ii. *One popular method is to obtain pseudo-label by clustering and use them to optimize the model.*
iii. *Although this kind of approach has shown promising accuracy, it is hampered by 1) noisy labels produced by clustering and 2) feature variations caused by camera shift. The former will lead to incorrect optimization and thus hinders the model accuracy.*
iv. *The latter will result in assigning the intra-class samples of different cameras to different pseudo-label, making the model sensitive to camera variations. In this paper, we propose a unified framework to solve both problems.*
v. *Concretely, we propose a Dynamic and Symmetric Cross-Entropy loss (DSCE) to deal with noisy samples and a camera-aware meta-learning algorithm (MetaCam) to adapt camera shift. DSCE can alleviate the negative effects of noisy samples and accommodate the change of clusters after each clustering step. MetaCam simulates cross-camera constraint by splitting the training data into meta-train and meta-test based on camera IDs.*
vi. *With the interacted gradient from meta-train and meta-test, the model is enforced to learn camera-invariant features. Extensive experiments on three re-ID benchmarks show the effectiveness and the complementary of the proposed DSCE and MetaCam. Our method outperforms the state-of-the-art methods on both fully unsupervised re-ID and unsupervised domain adaptive re-ID.*

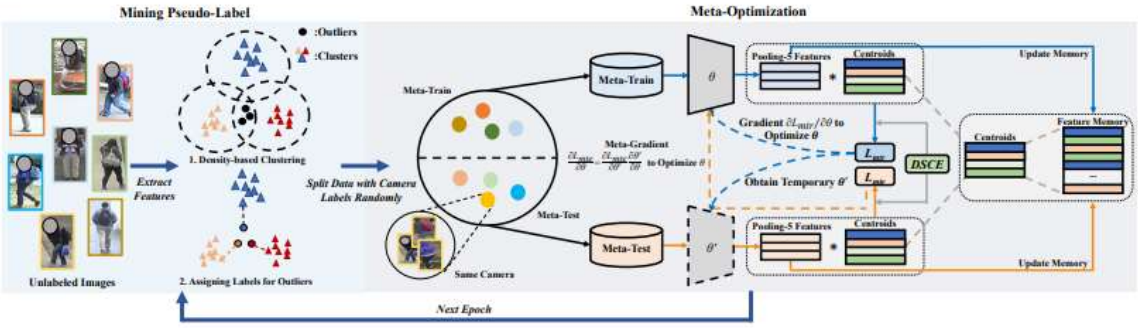


Fig: The framework for unsupervised re-ID, which includes two training stages, i.e., "Mining Pseudo-Label" stage and "Met Optimization".

The first stage assigns samples with pseudo-label based on DBSCAN.

The second stage splits the training data into meta-train and meta-test sets based on camera labels and optimizes the model with the proposed meta-learning strategy.

This camera-aware meta-learning (MetaCam) encourages the model to learn camera-invariant features.

To reduce the negative impact of noisy labels, we also propose a dynamic and symmetric cross-entropy loss (DSCE) that is used for both meta-train and meta-test data.

In our framework, the memory module saves the features of all samples, which enables us to dynamically build class centroids and thus to be adaptable to the change of clusters.

Watching You: Global-guided Reciprocal Learning for Video-based Person Re-identification(2021)

i. *Video-based person re-identification (Re-ID) aims to automatically retrieve video sequences of the same person under non-overlapping cameras.*
ii. *To achieve this goal, it is the key to fully utilize abundant spatial and temporal cues in videos. Existing methods usually focus on the most conspicuous image regions, thus they may easily miss out fine-grained clues due to the person varieties in image sequences.*
iii. *To address above issues, in this paper, we propose a novel Global-guided Reciprocal Learning (GRL) framework for video-based person Re-ID.*
iv. *Specifically, we first propose a Global-guided Correlation Estimation (GCE) to generate feature correlation maps of local features and global features, which help to localize the high- and lowcorrelation regions for identifying the same person. After that, the discriminative features are disentangled into high-correlation features and low-correlation features under the guidance of the global representations.*
v. *Moreover, a novel Temporal Reciprocal Learning (TRL) mechanism is designed to sequentially enhance the high-correlation semantic information and accumulate the low-correlation sub-*

critical clues. Extensive experiments are conducted on three public benchmarks. The experimental results indicate that our approach can achieve better performance than other state-of-the-art approaches.

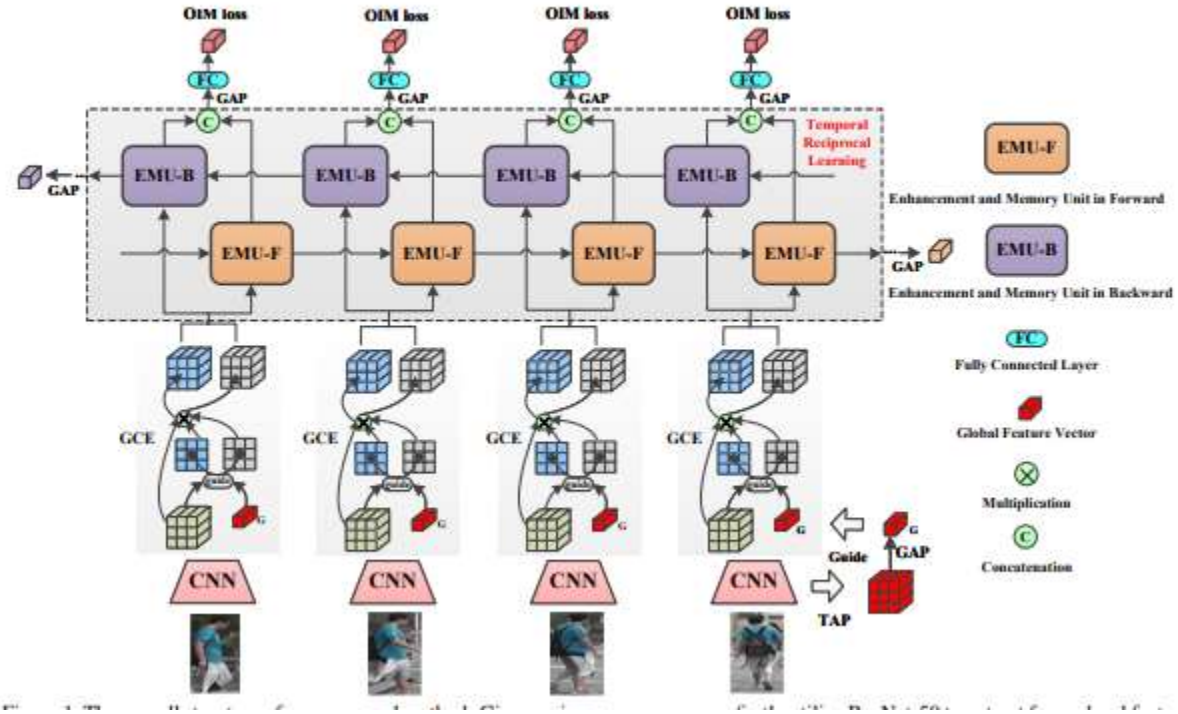


Fig: The overall structure of our proposed method.

Given an image sequence, we firstly utilize ResNet-50 to extract frame-level feature maps. Then, frame-level features are aggregated by TAP and GAP to generate a video-level feature. With the guidance of video-level features, Global Correlation Estimation (GCE) is utilized to generate the correlation maps for disentanglement. Afterwards, the Temporal Reciprocating Learning (TRL) is introduced to enhance and accumulate disentangled features in forward and backward directions.

Discover Cross-Modality Nuances for Visible-Infrared Person Re-Identification(2021)

I. *Visible-infrared person re-identification (Re-ID) aims to match the pedestrian images of the same identity from different modalities.*
II. *Existing works mainly focus on alleviating the modality discrepancy by aligning the distributions of features from different modalities.*
III. *However, nuanced but discriminative information, such as glasses, shoes, and the length of clothes, has not been fully explored, especially in the infrared modality.*

IV. *Without discovering nuances, it is challenging to match pedestrians across modalities using modality alignment solely, which inevitably reduces feature distinctiveness.*

V. *In this paper, we propose a joint Modality and Pattern Alignment Network (MPANet) to discover cross-modality nuances in different patterns for visibleinfrared person Re-ID, which introduces a modality alleviation module and a pattern alignment module to jointly extract discriminative features.*

VI. *Specifically, we first propose a modality alleviation module to dislodge the modality information from the extracted feature maps.*

VII. *Then, We devise a pattern alignment module, which generates multiple pattern maps for the diverse patterns of a person, to discover nuances.*

VIII. *Finally, we introduce a mutual mean learning fashion to alleviate the modality discrepancy and propose a center cluster loss to guide both identity learning and nuances discovering. Extensive experiments on the public SYSU-MM01 and RegDB datasets demonstrate the superiority of MPANet over state-of-the-arts.*

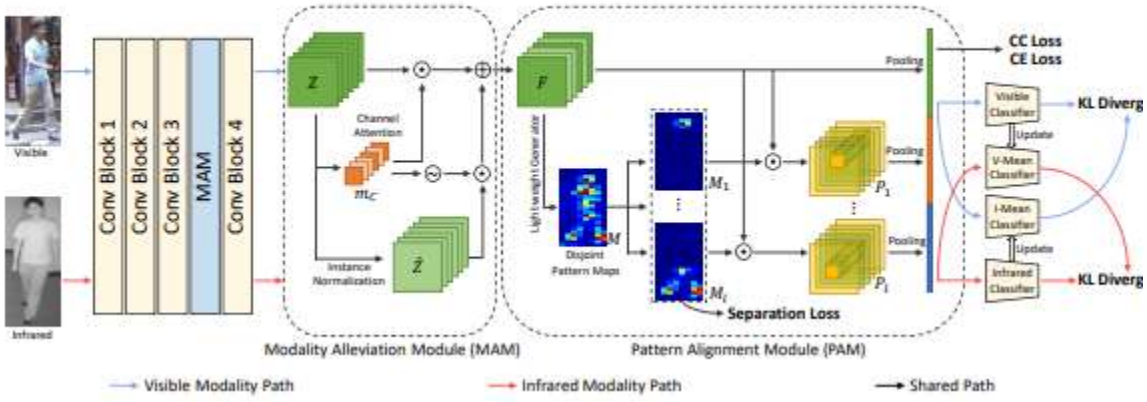


Fig : Framework of the proposed joint Modality and Pattern Alignment Network (MPANet).

The Modality Alleviation Module (MAM) receives feature maps from the former block to extract modality-irrelevant feature maps. Subsequently, the Pattern Alignment Module (PAM) generates pattern maps to discover nuances in different patterns. A separation loss is proposed to ensure the pattern maps focus on the different patterns. Then the proposed center cluster loss instructs each pattern map to focus on a certain pattern and guide identity learning with the cross-entropy loss jointly. For guiding the network to alleviate the modality discrepancy, two modality-specific classifiers are applied with two corresponding mean classifiers in a mutual mean learning fashion.

Matching on Sets: Conquer Occluded Person Re-Identification Without Alignment(2021)

I. *Occluded person re-identification (re-ID) is a challenging task as different human parts may become invisible in cluttered scenes, making it hard to match person images of different identities.*

II. *Most existing methods address this challenge by aligning spatial features of body parts according to semantic information (e.g. human poses) or feature similarities but this approach is complicated and sensitive to noises.*

III. *This paper presents Matching on Sets (MoS), a novel method that positions occluded person re-ID as a set matching task without requiring spatial alignment. MoS encodes a person image by a pattern set as represented by a 'global vector' with each element capturing one specific visual pattern, and it introduces Jaccard distance as a metric to compute the distance between pattern sets and measure image similarity.*

IV. *To enable Jaccard distance over continuous real numbers, we employ minimization and maximization to approximate the operations of intersection and union, respectively.*

V. *In addition, we design a Jaccard triplet loss that enhances the pattern discrimination and allows to embed set matching into deep neural networks for end-to-end training.*

VI. *In the inference stage, we introduce a conflict penalty mechanism that detects mutually exclusive patterns in the pattern union of image pairs and decreases their similarities accordingly.*

VII. *Extensive experiments over three widely used datasets (Market1501, DukeMTMC and Occluded-DukeMTMC) show that MoS achieves superior re-ID performance. Additionally, it is tolerant of occlusions and outperforms the state-of-the-art by large marginsfor Occluded-DukeMTMC.*

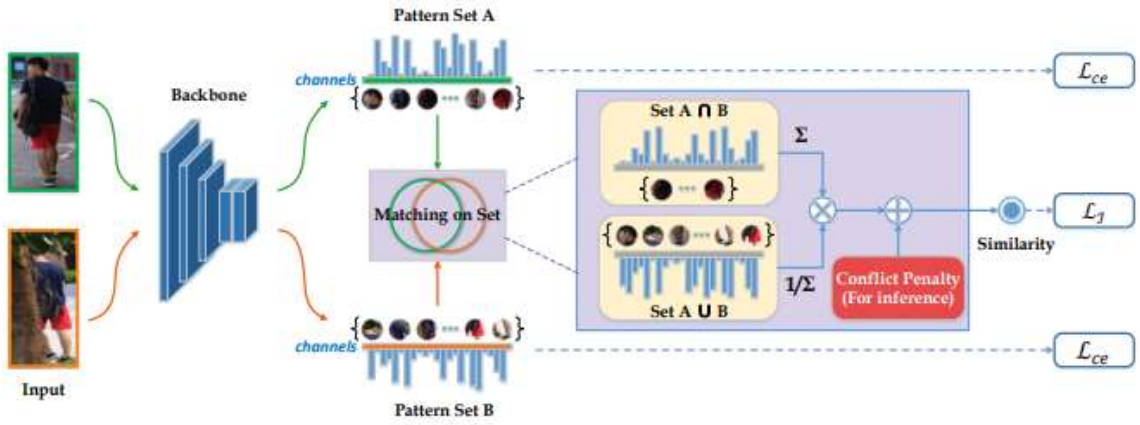


Fig : The framework of our proposed MoS:

Given a pair of person images as Input, MoS first encodes them by two pattern sets (Pattern Set A and Pattern Set B) where each set is represented by a 'global vector' with each element capturing one specific visual pattern.

It employs Jaccard similarity coefficient with a conflict penalty term as a metric to compute the distance between pattern sets and measure the image Similarity, more details to be described in Proposed Method.

Person30K: A Dual-Meta Generalization Network for Person Re-Identification(2021)

I. *Recently, person re-identification (ReID) has vastly benefited from the surging waves of data-driven methods. However, these methods are still not reliable enough for real world deployments, due to the insufficient generalization capability of the models learned on existing benchmarks that have limitations in multiple aspects, including limited data scale, capture condition variations, and appearance diversities.*

II. *To this end, we collect a new dataset named Person30K with the following distinct features: 1) a very large scale containing 1.38 million images of 30K identities, 2) a large capture system containing 6,497 cameras deployed at 89 different sites, 3) abundant sample diversities including varied backgrounds and diverse person poses.*

III. *Furthermore, we propose a domain generalization ReID method, dual-meta generalization network (DMG-Net), to exploit the merits of meta-learning in both the training procedure and the metric space learning.*

IV. *Concretely, we design a "learning then generalization evaluation" meta training procedure and a meta-discrimination loss to enhance model generalization and discrimination capabilities. Comprehensive experiments validate the effectiveness of our DMG-Net.*

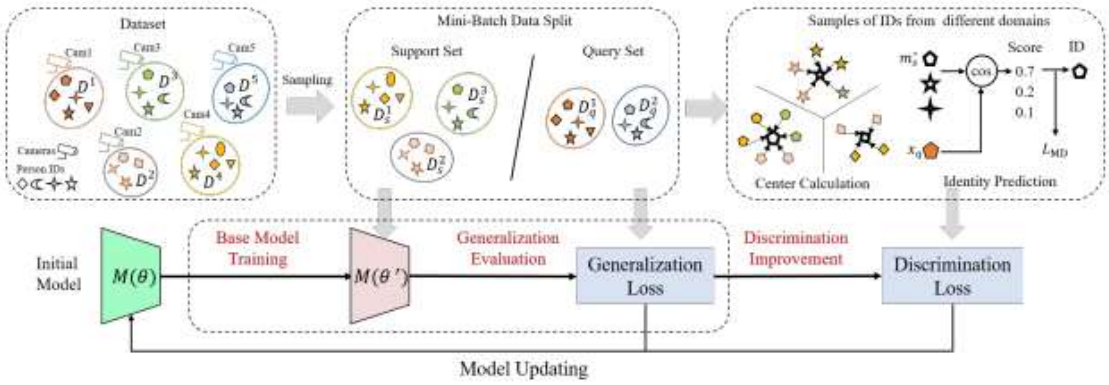


Fig: Illustration of our proposed DMG-Net.

DMG-Net includes a meta-generalization training procedure and a meta-discrimination loss.

The meta-generalization first trains a base model on the support set, then performs generalization evaluation on the query set.

For meta-discrimination loss, it optimizes the metric space to improve model discrimination

Learning 3D Shape Feature for Texture-insensitive Person Re-identification(2021)

It is well acknowledged that person re-identification (person ReID) highly relies on visual texture information like clothing. Despite significant progress has been made in recent years, texture-confusing situations like clothing changing and persons wearing the same clothes receive little attention from most existing ReID methods.

In this paper, rather than relying on texture based information, we propose to improve the robustness of person ReID against clothing texture by exploiting the information of a person's 3D shape. Existing shape learning schemas for person ReID either ignore the 3D information of a person, or require extra physical devices to collect 3D source data.

Differently, we propose a novel ReID learning framework that directly extracts a texture-insensitive 3D shape embedding from a 2D image by adding 3D body reconstruction as an auxiliary task and regularization, called 3D Shape Learning (3DSL). The 3D reconstruction-based regularization forces the ReID model to decouple the 3D shape information from the visual texture, and acquire discriminative 3D shape ReID features.

To solve the problem of lacking 3Dground truth, we design an adversarial self-supervised projection (ASSP) model, performing 3D reconstruction without ground truth. Extensive experiments on common ReID datasets and texture-confusing datasets validate the effectiveness of our mode.

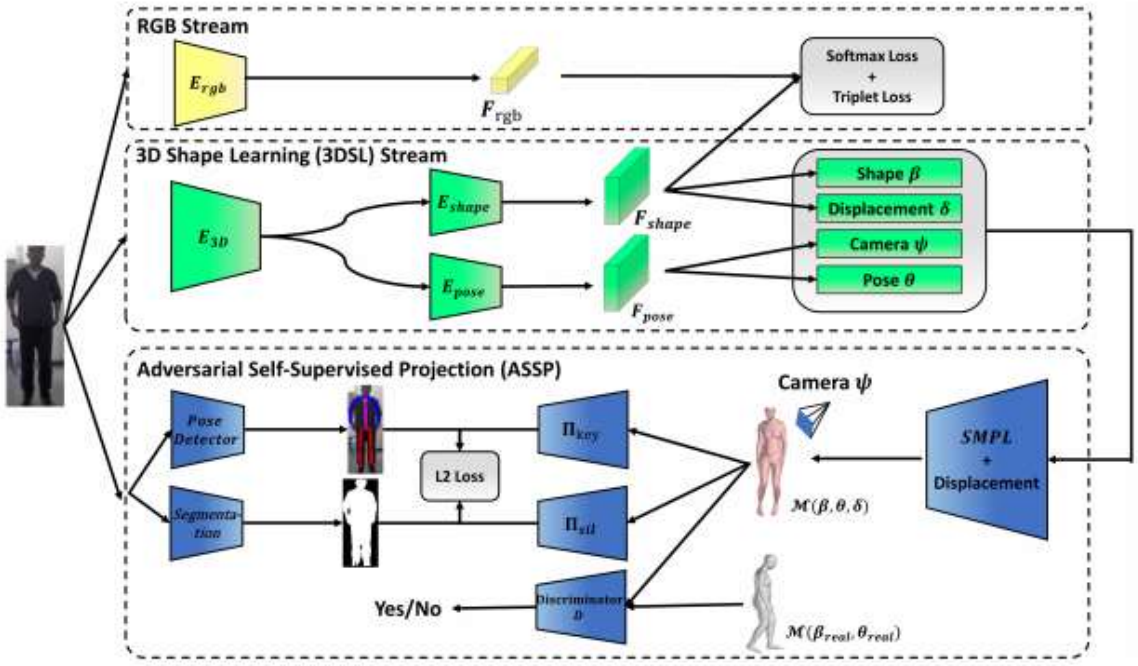


Fig: 2. The overview of the proposed model

The model consists of two branches. One is for learning 3D shape features under the regularization of 3D human reconstruction, named 3D Shape Learning.

Another branch is for learning texture insensitive RGB features from original images via metric learning,

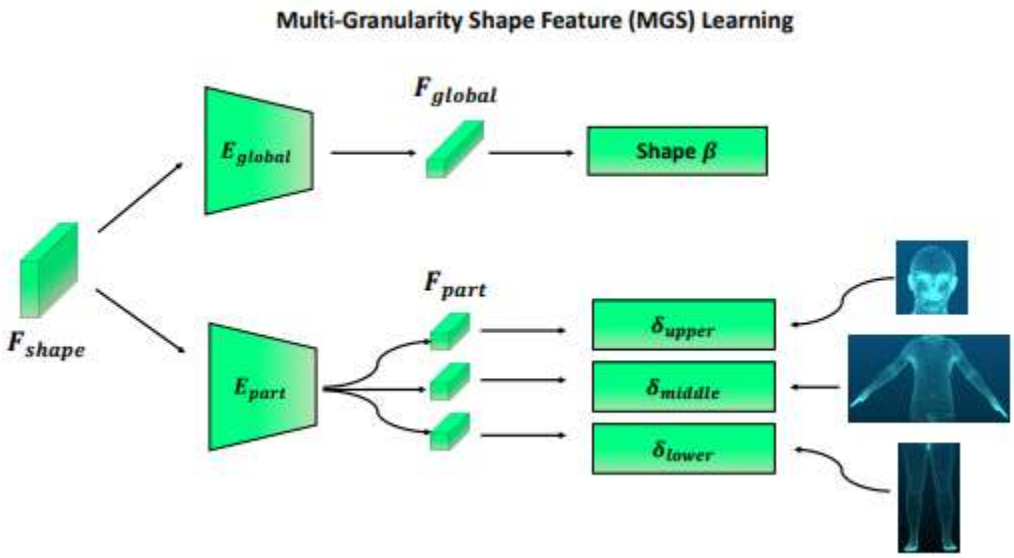


Fig: The illustration of Multi-Granularity Shape (MGS) learning.

In MGS, there is a global shape feature F global to estimate the global shape parameter β and part shape features Fpart to estimate part displacements.

We partition the 3D vertices into typically 3 parts. Eglobal and Epart both consists of shallow convolution and fully-connected layers.

Keynotes

To solve the problem of lacking ground truth to train 3D reconstruction, we introduce an unsupervised module called Adversarial Self-Supervised Projection (ASSP) to ensure coarse body manifolds via adversarial learning and fit the fine body details via self-supervised projection from 3D to 2D.

To enhance the discriminative ability of 3D shape features, we propose Multi-Granularity Shape (MGS) learning to capture part 3D shapes and increase the feature diversity

Partial Person Re-identification with Part-Part Correspondence Learning(2021)

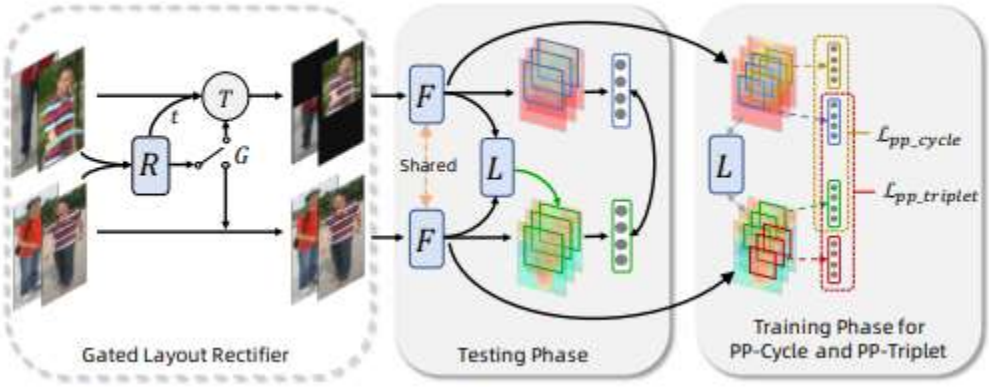
Driven by the success of deep learning, the last decade has seen rapid advances in person re-identification (re-ID).

Nonetheless, most of approaches assume that the input is given with the fulfillment of expectations, while imperfect input remains rarely explored to date, which is a non-trivial problem since directly apply existing methods without adjustment can cause significant performance degradation.

In this paper, we focus on recognizing partial (flawed) input with the assistance of proposed Part-Part Correspondence Learning (PPCL), a self-supervised learning framework that learns correspondence between image patches without any additional part-level supervision.

Accordingly, we propose Part-Part Cycle (PP-Cycle) constraint and Part-Part Triplet (PP-Triplet) constraint that exploit the duality and uniqueness between corresponding image patches respectively.

We verify our proposed PPCL on several partial person re-ID benchmarks. Experimental results demonstrate that our approach can surpass previous methods in terms of the standard evaluation metric.



The proposed PPCL framework mainly comprises a gated layout rectifier including R, G, T (GLRec), a backbone network F and a corresponding region locator L (CRLoc).

The GLRec module is a gated transformation regression CNN module that takes an arbitrary partial image in, and outputs a rectified result.

Then after feature extraction by F, the CRLoc module is employed to learning correspondence for part-part matching. Both GLRec and CRLoc modules are trained in a self-supervised manner to obtain the corresponding patches between two images without any part-level supervision.

Enhancing Diversity in Teacher-Student Networks via Asymmetric branches for Unsupervised Person Re-identification(2021)

i. *The objective of unsupervised person re-identification (Re-ID) is to learn discriminative features without labor intensive identity annotations.*
ii. *State-of-the-art unsupervised Re-ID methods assign pseudo labels to unlabeled images in the target domain and learn from these noisy pseudo labels.*
iii. *Recently introduced Mean Teacher Model is a promising way to mitigate the label noise. However, during the training, self-ensembled teacher-student networks quickly converge to a consensus which leads to a local minimum. We explore the possibility of using an asymmetric structure inside neural network to address this problem.*
iv. *First, asymmetric branches are proposed to extract features in different manners, which enhances the feature diversity in appearance signatures.*
v. *Then, our proposed cross-branch supervision allows one branch to get supervision from the other branch, which transfers distinct knowledge and enhances the weight diversity between teacher and student networks.*
vi. *Extensive experiments show that our proposed method can significantly surpass the performance of previous work on both unsupervised domain adaptation and fully unsupervised Re-ID tasks.*

Spatial-Temporal Correlation and Topology Learning for Person Re-Identification in Videos(2021)

a) *Video-based person re-identification aims to match pedestrians from video sequences across non-overlapping camera views.*
b) *The key factor for video person re-identification is to effectively exploit both spatial and temporal clues from video sequences.*
c) *In this work, we propose a novel Spatial-Temporal Correlation and Topology Learning framework (CTL) to pursue discriminative and robust representation by modeling cross-scale spatial-temporal correlation.*
d) *Specifically, CTL utilizes a CNN backbone and a key-points estimator to extract semantic local features from human body at multiple granularities as graph nodes. It explores a context-reinforced topology to construct multi-scale graphs by considering both global contextual information and physical connections of human body.*
e) *Moreover, a 3D graph convolution and a cross-scale graph convolution are designed, which facilitate direct cross-spacetime and cross-scale information propagation for capturing hierarchical spatial-temporal dependencies and structural information.*
f) *By jointly performing the two convolutions, CTL effectively mines comprehensive clues that are complementary with appearance information to enhance representational capacity.*
g) *Extensive experiments on two video benchmarks have demonstrated the effectiveness of the proposed method and the state-of-the-art performance.*

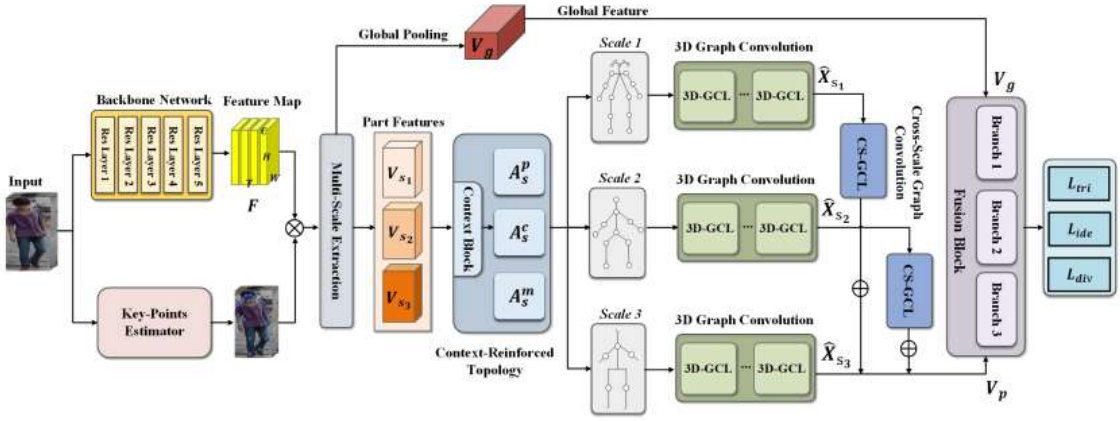


Figure: The overall architecture of the proposed CTL.

It consists of a backbone network with a key-points estimator, a context block, multiple 3D graph convolution layers (3D-GCLs), multiple cross-scale graph convolution layers (CS-GCLs) and a fusion block.

Unsupervised Pre-Training for Person Re-Identification(2021)

In this paper, we present a large scale unlabeled person re-identification (Re-ID) dataset "LUPerson" and make the first attempt of performing unsupervised pre-training for improving the generalization ability of the learned person Re-ID feature representation.

This is to address the problem that all existing person Re-ID datasets are all of limited scale due to the costly effort required for data annotation.

Previous research tries to leverage models pre-trained on ImageNet to mitigate the shortage of person Re-ID data but suffers from the large domain gap between ImageNet and person Re-ID data.

LUPerson is an unlabeled dataset of 4M images of over 200K identities, which is 30xlarger than the largest existing Re-ID dataset. It also covers a much diverse range of capturing environments (e.g., camera settings, scenes, etc.).

Based on this dataset, we systematically study the key factors for learning Re-ID features from two perspectives: data augmentation and contrastive loss. Unsupervised pre-training performed on this large-scale dataset effectively leads to a generic Re-ID feature that can benefit all existing person Re-ID methods.

Using our pre-trained model in some basic frameworks, our methods achieve state-of-the-art results without bells and whistles on four widely used Re-ID datasets: CUHK03, Market1501, DukeMTMC, and MSMT17.

Our results also show that the performance improvement is more significant on small-scale target datasets or under few-shot setting.

Person Re-identification using Heterogeneous Local Graph Attention Networks(2021)

I. *Recently, some methods have focused on learning local relation among parts of pedestrian images for person reidentification (Re-ID), as it offers powerful representation capabilities. However, they only provide the intra-local relation among parts within single pedestrian image and ignore the inter-local relation among parts from different images, which results in incomplete local relation information.*
II. *In this paper, we propose a novel deep graph model named Heterogeneous Local Graph Attention Networks (HLGAT) to model the inter-local relation and the intra-local relation in the completed local graph, simultaneously.*
III. *Specifically, we first construct the completed local graph using local features, and we resort to the attention mechanism to aggregate the local features in the learning process of inter-local relation and intra-local relation so as to emphasize the importance of different local features.*
IV. *As for the inter-local relation, we propose the attention regularization loss to constrain the attention weights based on the identities of local features in order to describe the inter-local relation accurately.*
V. *As for the intra-local relation, we propose to inject the contextual information into the attention weights to consider structure information.*
VI. *Extensive experiments on Market-1501, CUHK03, DukeMTMC-reID and MSMT17 demonstrate that the proposed HLGAT outperforms the state-of-the-art methods.*

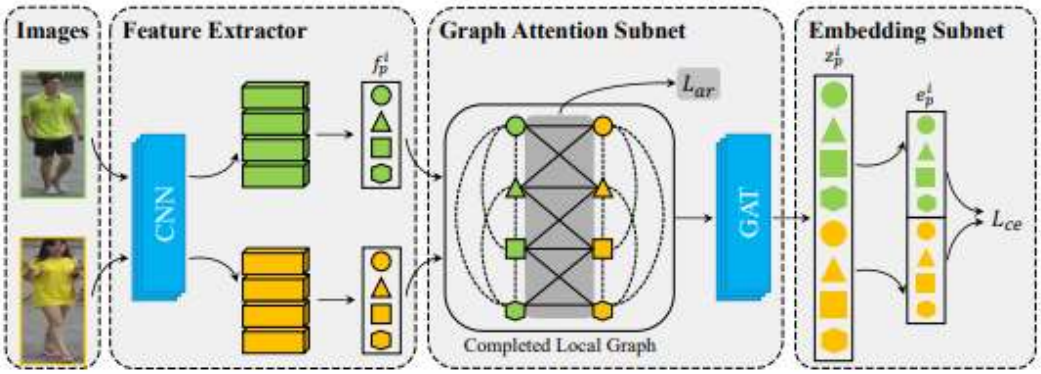


Figure: Pipeline of the proposed HLGAT. As for the completed local graph, the solid lines indicate the inter-local edges, and the dotted lines represent the intra-local edges

Overview

Feature Extractor. *Feed pedestrian images into CNN to obtain feature maps, and then we adopt uniform partition strategy to split these feature maps into several horizontal grids. Finally, we extract local features using the global max pooling operation on these grids.*

Graph Attention Subnet. *We regard the local features as the nodes to construct the completed local graph to learn the inter-local relation and the intra-local relation, simultaneously. These nodes are linked by the inter-local edges and the intra-local edges. For each node in the graph, we weight its all neighbor nodes using the attention weights, and then aggregate them to obtain the two kinds of relations. Meanwhile, we constrain the attention weights of inter-local edges using the attention regularization loss to describe the interlocal relation accurately, and we inject the contextual information into the attention weights of intra-local edges to consider structure information.*

• ***Embedding Subnet***. *In this subnet, we apply independent fully connected (FC) layers to reduce the dimension of the features extracted from Graph Attention Subnet. We utilize these dimension-reduced features as the final features, and make identity prediction on them.*

Generalizable Person Re-identification with Relevance-aware Mixture of Experts(2021)

Domain generalizable (DG) person re-identification (ReID) is a challenging problem because we cannot access any unseen target domain data during training.

Almost all the existing DG ReID methods follow the same pipeline where they use a hybrid dataset from multiple source domains for training, and then directly apply the trained model to the unseen target domains for testing. These methods often neglect individual source domains' discriminative characteristics and their relevances w.r.t. the unseen target domains, though both of which can be leveraged to help themodel's generalization. To handle the above two issues, we propose a novel method called the relevance-aware mixture of experts (RaMoE), using an effective voting-based mixture mechanism to dynamically leverage source domains' diverse characteristics to improve the model's generalization.

Specifically, we propose a decorrelation loss to makem the source domain networks (experts) keep the diversity and discriminability of individual domains' characteristics.

Besides, we design a voting network to adaptively integrate all the experts' features into the more generalizable aggregated features with domain relevance.

Considering the target domains' invisibility during training, we propose a novel learning-to-learn algorithm combined with our relation alignment loss to update the voting network.

Extensive experiments demonstrate that our proposed RaMoE outperforms the state-of-the-art methods.

Image-to-Video Person Re-Identification by Reusing Cross-modal Embeddings(2018)

Image-to-video person re-identification identifies a target person by a probe image from quantities of pedestrian videos captured by non-overlapping cameras. Despite the great progress achieved, it's still challenging to match in the multimodal scenario, i.e. between image and video. Currently, state-of-the-art approaches mainly focus on the task-specific data, neglecting the extra information on the different but related tasks.

In this paper, we propose an end-to-end neural network framework for image-to-video person re-identification by leveraging cross-modal embeddings learned from extra information.

Cross-modal embeddings from image captioning and video captioning models are reused to help learned features be projected into a coordinated space, where similarity can be directly computed.

Besides, training steps from fixed model reuse approach are integrated into our framework, which can incorporate beneficial information and eventually make the target networks independent of existing models.

Apart from that, our proposed framework resorts to CNNs and LSTMs for extracting visual and spatiotemporal features, and combines the strengths of identification and verification model to improve the discriminative ability of the learned feature.

The experimental results demonstrate the effectiveness of our framework on narrowing down the gap between heterogeneous data and obtaining observable improvement in image-to-video person re-identification.

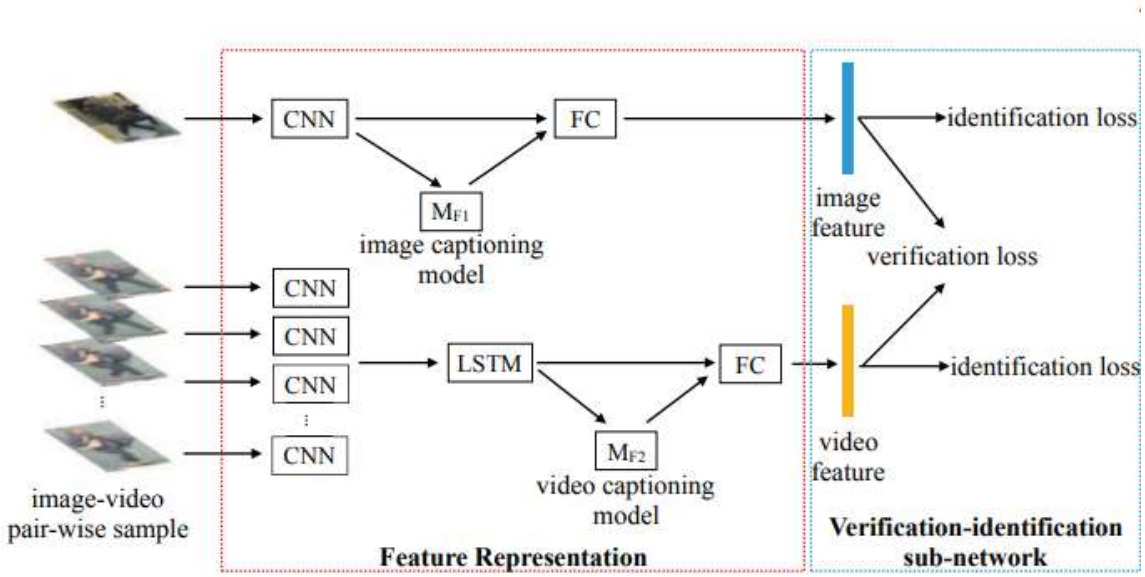


Fig: The pipeline of proposed framework for image-to-video person re-identification

On mainly in two parts,

feature representation*: used for extracting the feature representations of image and video,*

RNNs, particularly LSTMs, have shown promising stable and powerful performance in modeling long-range dependencies in sequence learning tasks. Therefore, we decide to combine CNNs and LSTM networks for learning the feature representations of image and video in this work

video person re-identification sub-network: *improving the discriminative ability of learned features and training a similarity metric between image and video modality, which can optimize the feature representation and similarity learning simultaneously.*

Mixed High-Order Attention Network for Person Re-Identification (2019)

Attention has become more attractive in person reidentification (ReID) as it is capable of biasing the allocation of available resources towards the most informative parts of an input signal. However, state-of-the-art works concentrate only on coarse or first-order attention design, e.g. spatial and channels attention, while rarely exploring higher-order attention mechanism. We take a step towards addressing this problem.

In this paper, we first propose the High-Order Attention (HOA) module to model and utilize the complex and high-order statistics information in attention mechanism, so as to capture the subtle differences among pedestrians and to produce the discriminative attention proposals.

Then, rethinking person ReID as a zero-shot learning problem, we propose the Mixed High-Order Attention Network (MHN) to further enhance the discrimination and richness of attention knowledge in an explicit manner.

Extensive experiments have been conducted to validate the superiority of our MHN for person ReID over a wide variety of state-of-the-art methods on three large-scale datasets, including Market-1501, DukeMTMC-ReID and CUHK03-NP.

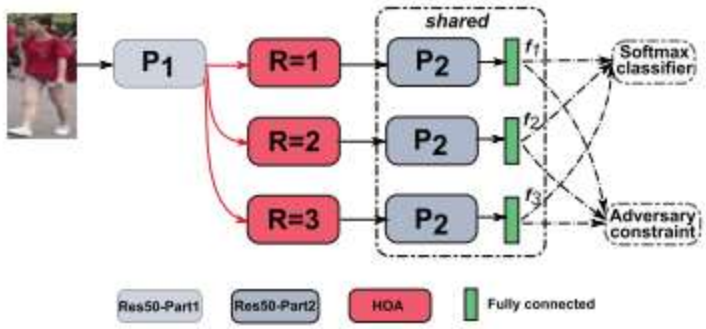


Fig: Illustration of Mixed High-Order Attention Network

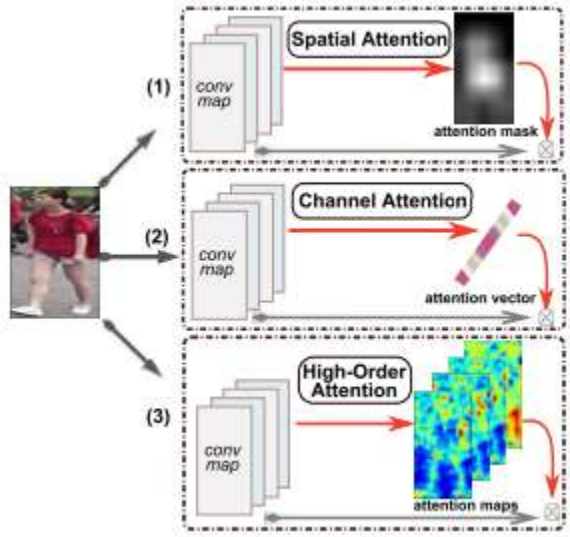


Fig: Attention Comparison

(1) ***Spatial attention*** *uses SoftMax-like gated functions to produce a spatial mask.*

(2) ***Channel attention*** *uses global average pooling and fully connected layers to produce a scale vector.*

(3) Our ***high-order attention*** *uses high-order polynomial predictor to produce scale maps that contain high-order statistics of convolutional activations.*

Notes

Zero-Shot Learning: In ZSL, the model is required to learn from the seen classes and then to be capable of utilizing the learned knowledge to distinguish the unseen classes.

It has been studied in image classification, video recognition and image retrieval/clustering. Interestingly, person ReID matches the setting of ZSL well where training identities have no intersection with testing identities, but most the existing ReID works ignore the problem of ZSL.

To this end, we propose Mixed High-Order Attention Network (MHN) to explicitly depress the problem of 'biased learning behavior of deep model 'caused by ZSL, allowing the learning of all-sided attention information which might be useful for unseen identities, preventing the learning of biased attention information that only benefits to the seen identities.

CUPR: Contrastive Unsupervised Learning for Person Re-identification

Most of the current person re-identification (Re-ID) algorithms require a large labeled training dataset to obtain better results.

For example, domain adaptation-based approaches rely heavily on limited real-world data to alleviate the problem of domain shift.

However, such assumptions are impractical and rarely hold, since the data is not freely accessible and require expensive annotation.

To address this problem, we propose a novel pure unsupervised learning approach using contrastive learning (CUPR).

Our framework is a simple iterative approach that learns strong high-level features from raw pixels using contrastive learning and then performs clustering to generate pseudo-labels.

We demonstrate that CUPR outperforms the unsupervised and semi-supervised state-of-the-art methods on Market-1501 and DukeMTMC-reID datasets/

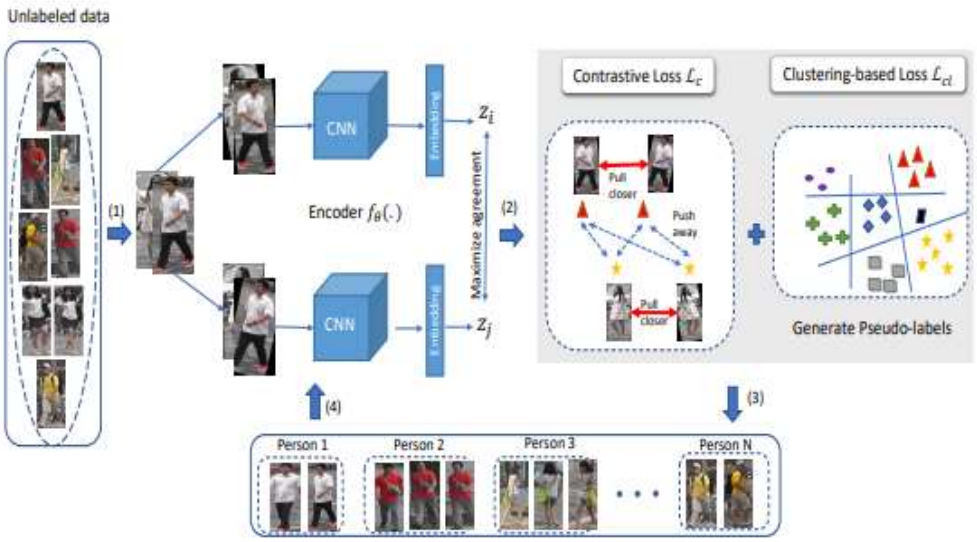


Fig : An illustration of the Contrastive Unsupervised Learning for Person Re-Identification

CUPR is an iterative framework mainly composed of a CNN backbone and aims to learn rich representations.

First, our framework extracts image features with a CNN model, then the HDBSCAN clustering method is performed using the feature similarities to generate pseudo-labels.

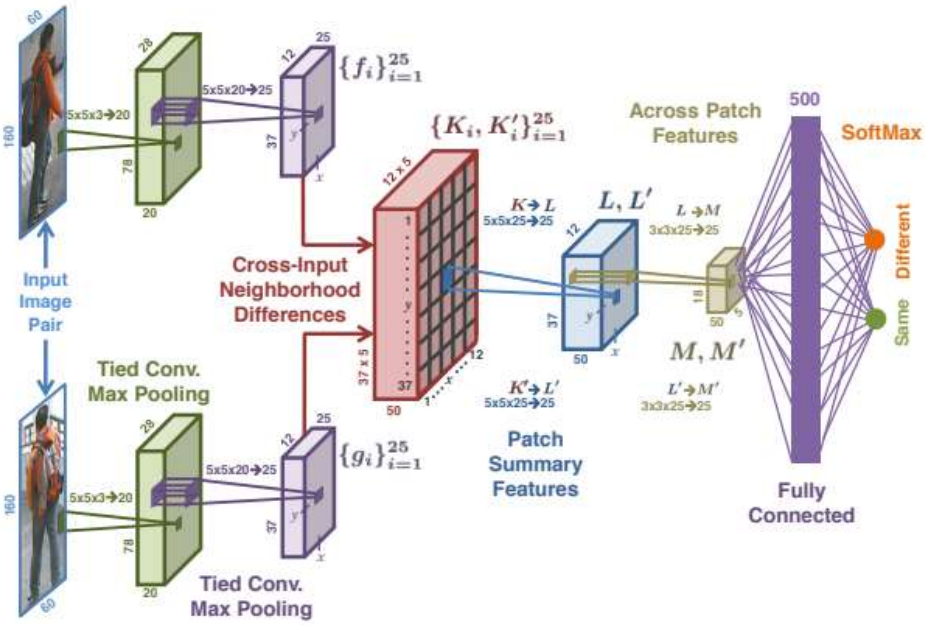
In an iterative end-to-end fashion, our model CUPR is trained using two optimization functions: (1) contrastive loss and (2) clustering-based loss.

An Improved Deep Learning Architecture for Person Re-Identification (2015)

In this paper, we propose a deep neural network architecture that formulates the problem of person re-identification as binary classification. Given an input pair of images, the task is to determine whether or not the two images represent the same person. our

network consists of the following distinct layers: two layers of tied convolution with max pooling, cross-input neighborhood differences, patch summary features, across-patch

features, higher-order relationships, and finally a softmax, function to yield the final estimate of whether the input images are of the same person or not


Input Image Pair
Tied Conv. Max Pooling
Tied Conv. Max Pooling
Cross-Input Neighborhood Differences
Patch Summary Features
Across Patch Features
Fully Connected
SoftMax
Different
Same
5x5x3→20
5x5x20→25
3x3x25→25
5x5x25→25